Федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский государственный университет аэрокосмического приборостроения»

На правах рукописи

Томчук Кирилл Константинович

Сегментация речевых сигналов для задач автоматической обработки речи

Специальность 05.12.13 – Системы, сети и устройства телекоммуникаций

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель –
кандидат технических наук,
с.н.с., доцент Корнеев Ю. А.

Санкт-Петербург – 2017

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Речевые технологии являются ключевым фактором в развитии автоматизированного окружения человека, начиная от совершенствования рабочих и исследовательских процессов и заканчивая областью персонального применения современных технологий. Работа подавляющего большинства речевых приложений невозможна без осуществления предварительной временной сегментации речи, то есть разделения речевого сигнала на квазистационарные по определенным характеристикам временные фрагменты.

В зависимости от стоящей перед конкретным речевым приложением задачи, применяемого метода решения и условий работы требуемый уровень сегментации речевого сигнала будет различаться. Это порождает большое многообразие частных задач сегментации и приводит к целесообразности разработки системных подходов к временной сегментации речевых сигналов.

Несмотря на высокую скорость развития вычислительной техники и информационных технологий, основные проблемы речевых приложений до сих пор остаются актуальными. Основной причиной является сложность структуры речевого сигнала: огромное разнообразие фонетических единиц языка, интонационных окрасок, личностных особенностей говорящего усугубляется разнообразием внешних факторов, влияющих на запись и передачу голоса. В результате речевые сигналы достаточно сложно детально исследовать и описывать с помощью математических моделей. Показательным является фактическое отсутствие систем распознавания русской речи со сверхбольшим словарем [1].

Перечисленные факторы определяют и основные недостатки существующих алгоритмов временной сегментации речевых сигналов: недостаточная точность определения границ сегментов, высокая ресурсоемкость, значительное ухудшение работы при наличии шумов.

Среди наиболее распространенных в мире языков нет ни одного, достаточно близкого русскому по генеалогической классификации языков, рассматривающей

общности языкового материала и языкового происхождения. Они не входят ни в восточную группу славянских языков, ни в сами славянские языки, ни в еще более крупную структуру – балто-славянскую языковую ветвь. Как следствие, фонетический состав и особенности произношения русского языка в значительной степени отличается от языков, для которых также активно разрабатываются речевые приложения, что затрудняет русскоязычную адаптацию языкозависимых зарубежных алгоритмов. Показательным является пример неудачного использования англоязычного ядра распознавания речи от мирового лидера рынка речевых технологий – компании Nuance Communications – в русскоязычной разработке [2].

Исходя из вышеизложенного, можно сделать вывод об актуальности создания новых и совершенствования имеющихся подходов к решению задачи временной сегментации речевых сигналов, и важности рассмотрения особенностей языка, на который данные алгоритмы ориентируются.

Степень разработанности темы. Фундаментальные труды по автоматической обработке речевых сигналов, во многом актуальные по сей день, принадлежат таким зарубежным и отечественным авторам, как Маркел Д. Д., Грэй А. Х., Рабинер Л. Р., Шафер Р. В., Фланаган Д. Л., Клатт Д., Фант Г., Винцюк Т. К., Косарев Ю. А. У истоков исследований, учитывающих специфику речевых сигналов русской речи, стоят отечественные ученые Златоустова Л. В., Потапова Р. К., Трунин-Донской В. Н., Бондарко Л. В., Вербицкая Л. А.; активное развитие русскоязычных речевых приложений прослеживается по работам современных российских исследователей, среди которых Сорокин В. Н., Галунов В. И., Кипяткова И. С., Мазуренко И. Л., Ронжин А. Л, Карпов А. А. и др.

Достаточно большое количество российских работ посвящено тематике сегментации речевых сигналов на различные уровни: Шарий Т. В., Жевуров С. В., Хлебников В. С., Петрушин В. А., Дорохин О. А., Старушко Д. Г., Федоров Е. Е., Шелепов В. Ю., Вишнякова О. А., Лавров Д. Н., Федоров В. М., Юрков П. Ю., Литвиненко С. Л., Ермоленко Т. В., Шевчук В. В., Галунов Г. В. и др. Однако лишь малая часть алгоритмов строится непосредственно в аспекте учета

особенностей русского языка: Конев А. А., Мещеряков Р. В., Бухаева О. Д., Сорокин В. Н., Цыплихин А. И., Аграновский А. В., Леднов Д. А. и др. Таким образом, внимание исследователей сосредоточено на определенных уровнях сегментации, в большинстве случаев – низких языконезависимых уровнях. Что актуализирует проведение системного анализа вопросов сегментации речевых сигналов с учетом применения их в первую очередь к русской речи.

Цели диссертационной работы – разработка алгоритмов автоматической многоуровневой временной сегментации речевых сигналов и вспомогательных алгоритмов.

Для достижения цели в диссертационной работе поставлены и решены следующие основные задачи:

1. Провести анализ:
    а. механизмов формирования звуков речи;
    б. спектра задач, возникающих при разработке алгоритмов сегментации речевых сигналов;
    в. существующих подходов к сегментации речевых сигналов.
2. Исследовать сигнальные особенности звуков русской речи:
    а. подготовить материал для исследования;
    б. разработать методику исследования;
    в. разработать исследовательское программное обеспечение
    г. получить и проанализировать статистические значения основных параметров звуков в зависимости от фонемы и положения в слове.
3. Разработать и апробировать алгоритмы сегментации:
    а. систематизировать спектр задач временной сегментации;
    б. разработать частные алгоритмы многоуровневой сегментации;
    в. разработать сопутствующие дополнительные алгоритмы.

Научная новизна состоит в следующем:

1. Разработана база данных для исследования сигнальных особенностей фонем с возможностью многокритериального извлечения статистических данных: по группе фонем, по диктору, по признаку ударности, по

положению фонем относительно границ слова, других фонем, ударного гласного.

2. Разработан алгоритм сегментации на периоды основного тона, использующий для анализа только отсчеты локальных экстремумов речевого сигнала.
3. Для увеличения эффективности MFCC-параметризации речевого сигнала на фоне шумов впервые предложено использовать психоакустическую модель одновременной слуховой маскировки и усиление сигнала на частотах кратных гармоник основного тона.
4. Предложен и апробирован подход к изменению темпа речи, основанный на модификации сегментов «пауза», «шумный», «взрывной», «вокализованный» речевого сигнала соответствующими подалгоритмами.

Теоретическая и практическая значимость работы.

1. Разработанный для исследования речевых сигналов программный комплекс:
   а. позволяет осуществлять автоматизированное транскрибирование русских слов;
   б. предоставляет интерфейс для первичной обработки РС;
   в. предоставляет интерфейс для ручной сегментации РС на произвольные типы сегментов и сохранения результатов в базу данных;
   г. осуществляет массовое вычисление сигнальных параметров для всех реализаций выбранной группы фонем.
2. Собрана информационная база значений основных параметров более чем 2000 вручную выделенных реализаций аллофонов с возможностью расширения как по количеству фонем, так и по количеству параметров.
3. Предложенная модификация алгоритма MFCC-параметризации позволяет получить относительное улучшение работы системы распознавания одиночных слов на 12% при усреднении по шумам в диапазоне ОСШ 0-20 дБ.

4. Разработанный алгоритм модификации темпа речи может быть использован как самостоятельное речевое приложение, имеющее, по результатам экспертных оценок, меньшее, чем у известных аналогов, количество артефактов звучания формируемого на выходе сигнала.

Методология и методы исследования. В исследовании использованы методы проектирования и анализа программных средств, общие методы системного анализа, методы теории вероятностей и математической статистики, цифровой обработки сигналов, спектрального анализа временных рядов, фонетики, психоакустики. Для проведения исследования применялось программирование в средах MATLAB, PHP, использована система управления базами данных MySQL.

Положения, выносимые на защиту. На защиту выносятся следующие положения и результаты:

1. Алгоритм сегментации речевого сигнала на периоды основного тона, основанный на фильтрации отсчетов локальных максимумов временной функции и позволяющий на порядок увеличить скорость сегментации и сохранить ее эффективность по сравнению с другими современными алгоритмами при ОСШ не менее 5 дБ.
2. Модифицированный алгоритм MFCC-параметризации, позволяющий за счет внедрения психоакустической модели частотного маскирования и усиления сигнала на частотах гармоник основного тона получить значительное улучшение работы системы распознавания одиночных слов на фоне шумов.
3. Алгоритм модификации темпа речевой фонограммы, использующий временную сегментацию для раздельной обработки типов речевой активности и пауз с собственными парциальными коэффициентами модификации.

Степень достоверности и апробация результатов. Разработанные алгоритмы обработки речевых сигналов и программные средства апробированы на обширном речевом материале, что отражено в тексте диссертационной работы.

Значительная часть разработанных алгоритмов сегментации речевых сигналов используется в компьютерной программе модификации темпа произнесения речи (НИР по гранту ПСП12377 правительства Санкт-Петербурга для студентов, аспирантов вузов и академических институтов, расположенных на территории Санкт-Петербурга, 2012 г.; НИР по гранту МК-4934.2012.9 Президента Российской Федерации, 2012-2013 г.; НИР ПСР-3.1.2–11 по целевой программе стратегического развития образовательного, научного и инновационного потенциала Санкт-Петербургского государственного университета аэрокосмического приборостроения как инновационного исследовательского университета, 2012-2013 г.; свидетельство о регистрации электронного ресурса № 20862 от 17.04.2015, ВНТИЦ 50201550159).

Основные положения и результаты диссертационной работы докладывались и обсуждались на следующих научных конференциях:

1. Научная сессия ГУАП, посвященная Всемирному дню космонавтики (г. Санкт-Петербург, ежегодно с 2009 по 2015 годы).
2. 20-я межвузовская научно-техническая конференция «Военная радиоэлектроника: опыт использования и проблемы, подготовка специалистов», посвященная 150-й годовщине со дня рождения А.С.Попова (г. Санкт-Петербург, 2009 г.)
3. Международная научная конференция «Системы и модели в информационном мире (СМИ-2009)» (г. Таганрог, 2009 г.)
4. Международная научная конференция «Современные исследовательские и образовательные технологии (СИОТ-2010)» (г. Таганрог, 2010).
5. Всероссийская научная конференция «Перспективы развития гуманитарных и технических систем» (г. Таганрог, 2011).

Личный вклад. Автором лично выполнены все этапы диссертационного исследования: постановка задач, подготовка исследовательской базы, создание методического, алгоритмического и программного обеспечения, проведение экспериментальных исследований, обработка и интерпретация данных, формулировка выводов.

Публикации. По теме диссертации опубликовано 15 печатных работ, в том числе три статьи в рецензируемых журналах из списка ВАК РФ. Получено свидетельство о регистрации электронного ресурса.

Объем и структура работы. Диссертация состоит из введения, четырех разделов, заключения, списка сокращений и условных обозначений, списка литературы и четырех приложений. Основной текст диссертационной работы изложен на 197 страницах, включает 86 рисунков, 18 таблиц, 4 приложения. Список литературы содержит 137 наименований.

1 ОБЩАЯ ПРОБЛЕМА АНАЛИЗА И СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ

Обработка речевых сигналов лежит в основе широкого спектра технических задач. Однако ввиду сложности структуры речевого сигнала, недостаточной изученности механизмов как речеобразования, так и речевосприятия, достигнутый уровень решения речевых задач в значительной степени не достигает уровня, с которым эти задачи решаются организмом человека. Анализ существующих подходов к обработке речевых сигналов и отдельных фрагментов речевых сигналов в различных речевых приложениях позволяет определить круг неразрешенных проблем, препятствующих созданию высокоэффективных технических решений данных задач.

1.1 Проблематика задачи автоматической сегментации речевых сигналов

1.1.1 Речевые технологии: актуальность, уровень развития

Речевые технологии позволяют создавать интуитивно понятные, легкодоступные и быстрые в применении интерфейсы для «общения» человека с компьютеризированным техническим окружением.

Перечень актуальных приложений, которые могут быть реализованы за счет речевых технологий, крайне обширен. Это может быть голосовой помощник для электронных и аудиокниг; детектор телефонных номеров и электронных адресов (e-mail), произнесенных в голосовых сообщениях; мастер протоколирования совещаний, в том числе идентификацией текущего оратора; голосовой поиск информации; голосовая навигация (синтез и распознавание) и т. д. [3]. Основные классы задач систем обработки речи приведены на рисунке 1.1.

Рисунок 1.1 – Задачи, решаемые речевыми приложениями

Особую роль в речевых технологиях играет область, относящаяся автоматическому к распознаванию и восприятию человеческой речи. Активные исследования в области распознавания речи начались около 60 лет назад. Работы велись в таких организациях, как Bell Laboratories, RCA Labs, University College в Англии, MIT Lincoln Labs, НИИ Дальней связи (г. Ленинград), Институт проблем передачи информации РАН [3]. Первое устройство для автоматического распознавания речи появилось в 1952 году и было предназначено для автоматического распознавания раздельно произносимых цифр [4, 5].

Однако механизмы восприятия речи человеком до сих пор не изучены достаточно глубоко, так как на практике изучение процесса обработки человеческим мозгом получаемой информации является крайне сложной задачей. В итоге, архитектура существующих систем распознавания имеет мало общего с архитектурой человеческого восприятия речи [3].

Одной из первых работ, посвященных обработке речевых сигналов (РС), является монография американского ученого Джеймса Л. Фланагана [6]. В ней рассматриваются:

- процессы речеобразования, представления голосового тракта в виде различных моделей (электрической, аппроксимации трубами);
- вопросы акустического восприятия;
- вопросы анализа фонограмм: использование спектрального анализа (следует отметить использование кратковременного частотного анализа), формантного анализа, выделение частоты основного тона (ОТ),
- устройства сжатия речи (вокодеры).

Развитие методов цифровой обработки расширило возможности обработки речевых сигналов. Это видно в трудах известных зарубежных авторов: Рабинера Л. Р. и Голда Б. [7], Рабинера Л. Р. и Шафера Р. В. [8], Оппенгейма А. [9], Маркела Дж. Д. и Грея А. Х. [10] – работы которых стали «классикой» в областях обработки речи и цифровой обработки сигналов.

Таким образом, круг возможных применений технологий автоматической обработки речи чрезвычайно обширен. Однако, несмотря на многочисленные исследования в течение последних 60 лет, многие даже основные задачи данной области так и не были полностью решены.

1.1.2 Применение сегментации речевых сигналов в речевых приложениях

Временная сегментация речевых сигналов является базовой задачей в любой голосовой системе и необходима для ее эффективной работы [11, 12, 13]. В зависимости от предназначения речевого приложения требуется различный уровень сегментации: для одних задач достаточно сегментации «речь/пауза»; для других может потребоваться сегментация на характерные речевые фрагменты (вокализованные, шумные, взрывные, паузы-смычки), например, для задач верификации диктора, модификации параметров речи; для иных приложений необходима сегментация до уровня фонем (например, распознавание речи), до отдельных периодов колебаний голосовых связок в огласованных звуках.

Одним из уровней сегментации является сегментация на широкие фонетические классы (ШФК, рисунок 1.2). Результаты сегментации на ШФК

могут быть использованы в задаче идентификации диктора [14]: голос диктора описывается множеством моделей, по одной на каждый ШФК. Данный вид сегментации может быть также положен в основу системы распознавания речи с малым словарем [15].

Кроме того, для решения задач идентификации и верификации диктора важны характерные признаки голоса, присущие определенным сегментам РС: значения высших формантных частот полостей речевого аппарата и частота колебаний голосовых связок определяются на вокализованных сегментах, при этом, зачастую, для выделения индивидуальных характеристик голоса следует рассматривать отдельные колебания [16, 17].

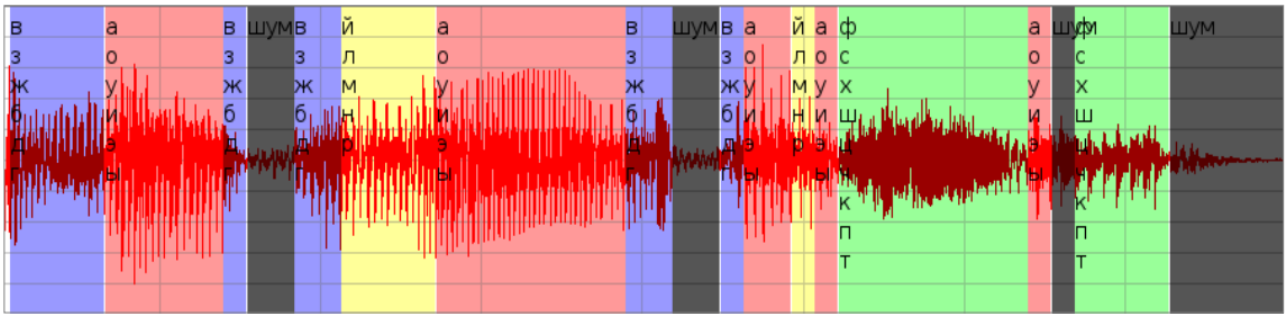


Рисунок 1.2 – Результат автоматической временной сегментации РС на ШФК [14]

В задачах сжатия РС высокую эффективность показывают вокодерные методы, использующие параметризацию для компактного представления РС [18]. Для работы класса полосных вокодеров требуется сегментация РС «тон/не тон», а также оценка частоты ОТ. Кроме того, в других типах вокодеров можно добиться более высоких коэффициентов сжатия путем применения различных алгоритмов для разных сегментов РС [19]: переходные участки речи с быстрой артикуляцией несут информацию одновременно и о предыдущем звуке, и о следующем, а длительные вокализованные фрагменты и паузы между словами, в свою очередь, имеют гораздо меньшую по времени плотность информации.

Вокодеры используются также в приложениях модификации характеристик речи (изменение скорости произнесения, тембра диктора) [8]. В то же время, в разделе 4 диссертационной работы представлен разработанный алгоритм модификации скорости произнесения речи, осуществляющий раздельную

обработку типов речевой активности, т. е. основанный на результатах глубокой временной сегментации: «речь/пауза», «шумный/вокализованный/взрывной», ОТ-сегментация. Как будет показано, такой подход дает предложенному алгоритму ряд значимых преимуществ перед существующими вокодерными методами.

В задаче модификации интонационных характеристик речи основную роль играют вокализованные звуки: изменение интонации осуществляется за счет изменения частотных характеристик квазипериодических колебаний ОТ РС, а также модулирующей функции последовательности таких колебаний [20, 21]. Таким образом, для решения данной задачи требуется сегментация «тон/не тон», а для вокализованных фрагментов («тон») – дополнительная сегментация на отдельные периоды ОТ с целью модификации их характеристик.

Аналогичный подход реализации интонационной окраски речи применим в задаче конкатенативного синтеза речи: сформированные последовательности фонем для придания необходимого звучания подвергаются модификации на уровнях изменения модулирующей функции и частотных характеристик колебаний ОТ, – что требует осуществления ОТ-сегментации РС [17].

Решение вопросов шумоочистки РС осложняется большим разнообразием типов возможных помех [22, 23]. Для устранения влияния стационарных на некоторых интервалах времени шумовых или периодических помех может адаптивно применяться режекторная фильтрация. Однако для устранения импульсных помех необходимы иные подходы. В частности, если импульсная помеха затрагивает вокализованный звук, она может быть устранена за счет замещения отдельных периодов ОТ на результат векторной интерполяции незатронутых помехой периодов, расположенных по обе стороны от нее.

Задачи автоматической временной сегментации РС можно разделить на два вида: сегментация при априорно известной последовательности фонем соответствующей фразы (контекстно-зависимая сегментация) [24, 25]; и сегментация при изначально отсутствующих данных об информационном содержании сигнала (контекстно-независимая сегментация) [26].

При реализации контекстно-зависимой сегментации основополагающей является операция автоматического транскрибирования текста, позволяющая учесть последовательность фонем, соответствующую данному РС [24, 27]. Такая сегментация, в сравнении с контекстно-независимой, показывает значительно лучшие результаты по точности разметки фонограмм.

В рамках диссертационной работы разработан алгоритм автоматизации транскрибирования русских слов, описание данного алгоритма представлено в Приложении А, подраздел А.2.

Задача автоматической контекстно-независимой сегментации (априорная информация о последовательности фонем в речи неизвестна) до сих пор полностью не решена [28, 29, 30].

Существует два подхода к решению задачи сегментации РС: разделение на фиксированные по длительности участки с последующим распознаванием их принадлежности к определенным группам / фонемам (см., например, [31, 32]); и фонемная сегментация, при которой РС делится на синтагмы вплоть до отдельных фонем. Из-за отсутствия надежных алгоритмов фонемной сегментации, в современных системах распознавания речи преобладает первый подход [33].

Таким образом, сегментация РС является неотъемлемой частью речевых приложений. При этом в зависимости от специфики реализации приложения, может использоваться сегментация разной глубины и на разные типы сегментов. Сегментация может быть контекстно-зависимая и контекстно-независимая, а по принципу определения границ сегментов существует сегментация на фрагменты фиксированной длительности и фонемная сегментация. Наиболее сложной задачей является контекстно-независимая фонемная сегментация.

1.1.3 Произнесение и восприятие речи человеком. Фонетическое строение сигнала русской речи

Если рассматривать строение речевого аппарата как акустической системы, то его удобно представить в виде трех функциональных блоков [35]:

– генератор: воздушный резервуар (легкие), мышечная система, выводной

канал (трахея и гортанная трубка);

– вибраторы: голосовые связки;
– резонаторы: глотка, ротовая и носовая полости – образуют т. н. артикуляционную систему.

Частотным диапазоном речи человека принято считать интервал от 500 до 2000 Гц [36]. В частности, подобный диапазон используется для передачи речи в системах телефонии; наиболее распространенным диапазоном при этом являются частоты от 280 Гц до 3,3 кГц [37]. Такие значения выбираются, в свою очередь, в соответствии с аудиограммой чувствительности слуха на разных частотах: нормально слышащий человек хорошо воспринимает частоты от 250 Гц до 8 кГц, наилучшая чувствительность слуха достигается на интервале от 500 Гц до 4 кГц согласно [38], от 500 Гц до 2 кГц согласно [39].

Основная частота колебаний голосовых связок (частота ОТ) находится в пределах от 50 до 250 Гц для мужчин и от 120 до 500 Гц для женщин [40]. Однако, несмотря на то, что частота ОТ оказывается за пределами нижней границы описанных выше диапазонов человеческой речи, восприятие речи не искажается: человеческое ухо компенсирует недостающую гармонику основного тона на основе гармоник кратных частот [39].

В качестве модели, хорошо описывающей артикуляцию речевого аппарата, используется авторегрессионная модель РС [41, 42]. Авторегрессионный процесс описывается разностным уравнением:

$$x(n) = \sum_{i=1}^{P} a(i)x(n-i) + \xi(n), \qquad (1.1)$$

где $x(n)$ – вектор отсчетов сигнала размером n; $a(i)$ – авторегрессионные коэффициенты процесса; P – порядок процесса; $\xi(n)$ – порождающий процесс.

Выбор порядка P модели зависит от требуемых качественных характеристик алгоритма, объема имеющихся данных и представляет собой оптимизационную задачу. Высокими динамическими характеристиками оценок параметров авторегрессионной модели отличается метод Берга [43].

Передаваемое в РС речевое сообщение может быть рассмотрено как последовательность фонем. В каждом языке выделяют обычно от 13–14 (некоторые языки Австралии и Океании) до 70 и более фонем (в кавказских языках). Обычно число фонем в языке близко к 36–40 [44]. В русском языке принято выделять 43 фонемы (37 согласных и 6 гласных) [44, 45, 46, 47], в украинском – 37 [44], в английском – 42 фонемы [8].

Сильные различия в строении языков, в их фонетическом составе значительно усложняют русскоязычную адаптацию иностранных алгоритмов обработки РС. Например, согласно Международному фонетическому алфавиту SAMPA, в американском английском языке 24 согласных и 17 гласных (против 37 согласных и 6 гласных в русском) – из-за такого соотношения гласных и согласных выделение фонем в русской речи является более трудной задачей, так как согласные звуки сложно распознавать из-за их большой вариабельности и маленькой длительности [48].

В технической литературе [49, 50, 14] можно встретить различные классификации фонем русской речи (рисунки 1.3, 1.4, 1.5).

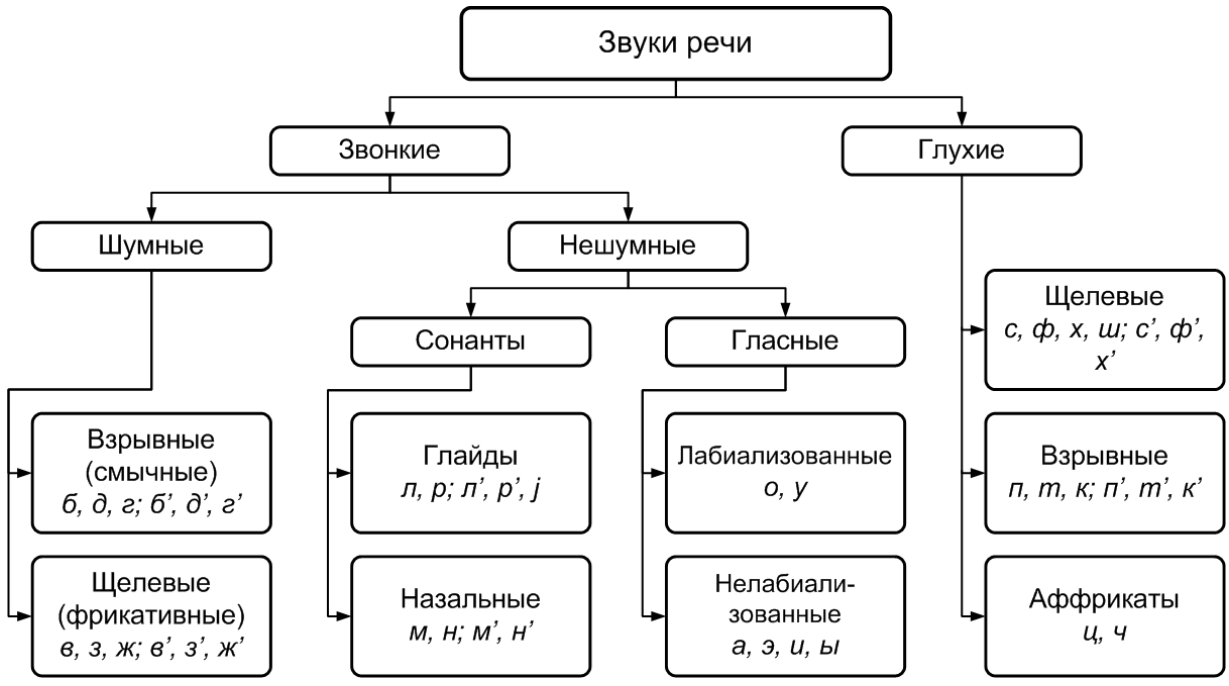


Рисунок 1.3 – Классификация фонем русского языка, приведенная в книге Косарева Ю. А. [49]

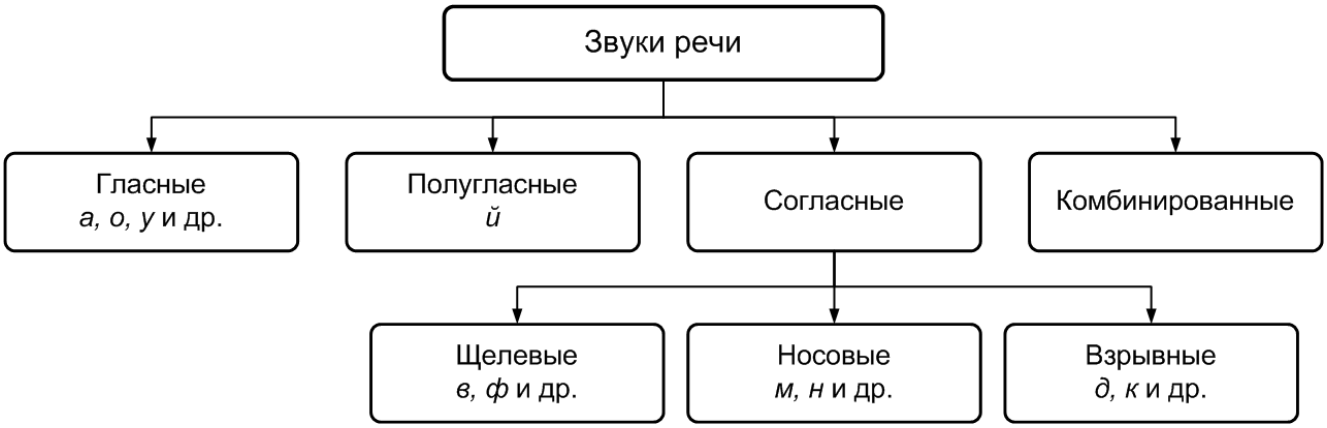


Рисунок 1.4 – Классификация звуков русской речи согласно Николенко Л. А. [50]

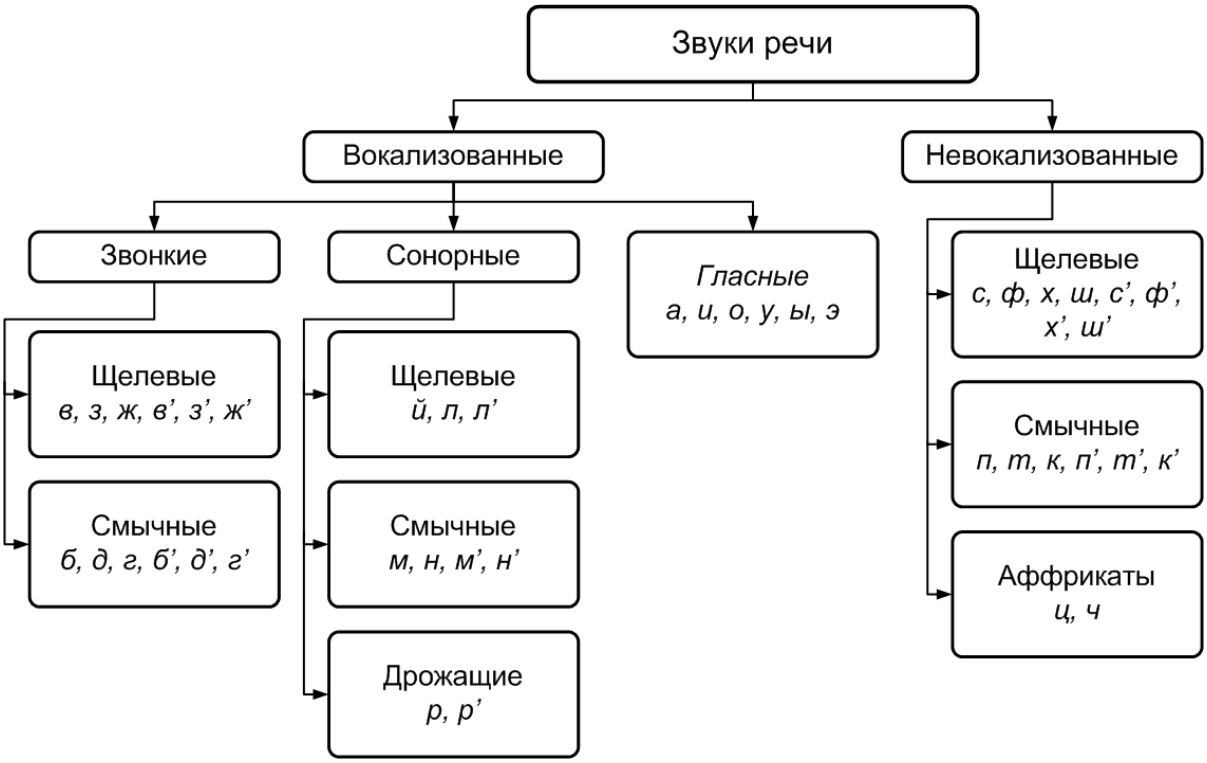


Рисунок 1.5 – Классификация звуков русской речи согласно Клименко Н. С. [14]

На рисунке 1.4, как пишет его автор, под комбинированными звуками понимаются дифтонги и аффрикаты (согласные звуки, характеризующиеся наличием полной преграды, которая затем переходит в щель; для русского языка это звуки [ц] и [ч], см. [14]). В данной классификации применение термина «дифтонг» требует специальной оговорки, так как в русском языке нет дифтонгов [51]. Вероятно, под дифтонгами автором подразумеваются неоднородные гласные, называемые дифтонгоидами. Дифтонгом, в свою очередь, называется

гласный, состоящий из двух отличающихся по звучанию, но близких по длительности и образующих один слог частей. К примеру, в английском языке существует 8 дифтонгов (например, в словах «go», «time») и 2 трифтонга (три гласных элемента, произносимых слитно в одном слоге). В русском же языке сочетания гласных звуков всегда принадлежат разным слогам. У дифтонгоида, в отличие от дифтонга, длительности составляющих его частей различаются значительно. К примеру, к дифтонгоидам можно отнести все гласные, кроме [и], произносящиеся в соседстве с мягкими согласными (например, ударный [^{и}а] в слове «пятый»).

Гласные звуки русского языка артикуляционно классифицируются по степени подъема языка и по его активной области (рисунок 1.6). С сигнальной точки зрения эти артикуляционные признаки обуславливают конфигурацию резонансных полостей речевого аппарата, которая отражается в значениях формантных частот вокализованных звуков. В частности, изменение первой формантной частоты *F1* свидетельствует об изменении открытости-закрытости, а изменение второй формантной частоты *F2* – об изменении ряда. Понижение частот *F1* и *F2* может быть обусловлено лабиализованностью звука [52]: звук [у] на рисунке 1.6.

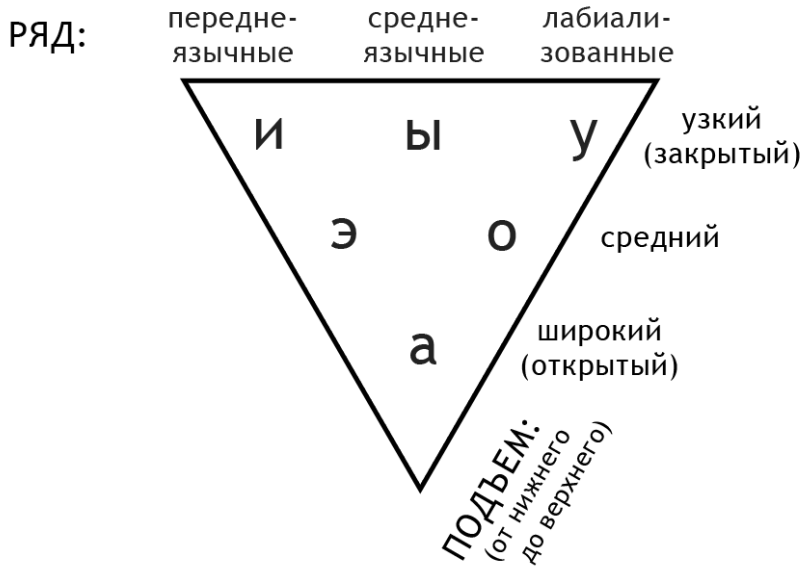


Рисунок 1.6 – Треугольник русских гласных звуков

Касательно особенностей согласных звуков, можно обратить внимание на спектр турбулентного шума, имеющего плоский участок в диапазоне от 500 до 3000 Гц, выше и ниже которого он спадает со скоростью 6 дБ на октаву [28].

Таким образом, русский язык представлен 37 согласными и 6 гласными фонемами. Произношение фонемы определяется текущей конфигурацией речевого аппарата говорящего. Несмотря на глубокую связь произносимых фонем с их сигнальными параметрами, в технической литературе отсутствует единство в вопросе структурной классификации фонем и их вариаций. В силу особенностей состава звуков русского языка адаптация языкозависимых алгоритмов, разработанных для других языков, является сложной задачей.

1.1.4 Параметризация сегментов речевого сигнала

Извлечение (вычисление) ряда параметров РС и дальнейшая их кластеризация являются неотъемлемой частью процесса сегментации. Следует отметить, что в данном пункте рассмотрены общие вопросы параметризации РС. Непосредственно основной ряд параметров, применяемых для автоматической временной сегментации, рассмотрен в подразделе 2.3 «Вычисление и анализ ряда сигнальных параметров реализаций фонем русского языка».

Речевой сигнал имеет сложную структуру и неустойчив сразу по многим параметрам: длительности фонем, темпа, высоты голоса; большую роль играют индивидуальные физиологические особенности, активная артикуляция, эмоциональное состояние человека. Это затрудняет применение в данной области методов анализа искусственных сигналов. Сложной задачей является выбор набора признаков, позволяющих достаточно полно и компактно описать РС [53].

В связи с различиями произношения статистические параметры по укрупненной классификации одних и тех же звуков (например, русского [т] и английского [t]) могут различаться для разных языков в связи с различиями артикуляции. В то же время, такие различия минимальны для языков восточнославянской группы (русский, украинский, белорусский). В работе [54] приводятся результаты исследования плотностей распределения формант по

частоте для русского и украинского языков. Делается вывод о высокой степени сходства вероятностных свойств русских и украинских речевых сигналов. В то же время отдельный акцент делается на заметное различие таких свойств для мужских и женских голосов.

Временные параметры звуков играют также большую роль в автоматическом синтезе речи, поэтому просодии РС и длительностям фонетических сегментов посвящено большое количество исследований. Например, в работах [55, 56] приводятся правила для синтеза длительности сегментов в различных контекстах, а в работах [57, 58] предпринимается попытка разработки алгоритмов управления длительностью сегментов в речевом потоке.

Известно, что на длительность звука влияет стиль произношения и присущая каждому диктору скорость артикуляции. Кроме того, длительность зависит от положения звука в слове и положения слова во фразе. Имеется более десяти параметров, влияющих на длительность звуков [59], что делает табличное описание всех возможных вариантов невозможным. Так, в произнесенной фразе звуки, находящиеся близко к ее началу, имеют значительно меньшую длительность по сравнению со звуками, лежащими ближе к ее концу. Существуют также звукосочетания «гласный–согласный», в которых, в зависимости от типа каждого из двух сегментов, изменяется их длительность друг относительно друга, в то время как суммарная их длительность практически постоянна.

Длительности пар соседних фонетических сегментов в подавляющем большинстве случаев сильно коррелированны [60].

Отдельную группу представляют параметры, относящиеся к вокализованным звукам. Во временной области вокализованные звуки характеризуются явно выраженной квазипериодичностью сигнала, связанной с колебаниями голосовых связок. Частота колебаний голосовых связок при произнесении звонких звуков называется частотой основного тона и является одной из индивидуальных особенностей человека, зависящей от длины голосовых связок, их массы и натяжения. Частота ОТ непостоянна в процессе речи и может заметно изменяться даже в пределах одного звука, в особенности у ударных

гласных. При этом длина и масса голосовых связок являются врожденными, а изменение частоты ОТ происходит за счет варьирования натяжения связок, на которое, помимо прочего, оказывает влияние громкость произнесения (при повышении громкости речи высота тона обычно растет) [61].

Другой важнейшей характеристикой вокализованных звуков (а также особенностей строения речевого аппарата конкретного человека) являются формантные частоты, которые представляют собой резонансные частоты акустических полостей речевого аппарата.

Резонансные частоты полости простой трубки постоянного сечения следуют через равные по частоте интервалы ΔF [35]:

$$\Delta F = (2n-1)\cdot\frac{c}{4L} \tag{1.2}$$

где n – целое положительное число, c – скорость звука в воздухе ~350 м/с, L – длина трубки, у взрослого человека длина речевого тракта (от голосовых связок до губ) составляет около 17 см.

Сечение голосового тракта не является постоянным, поэтому резонансные частоты, в отличие от описываемого формулой (1.2) случая, будут разнесены на разные частотные интервалы. У мужчин среднее расстояние между формантами, зависящее от длины голосового тракта, составляет около 1000 Гц [62]. У женских и детских голосов среднее расстояние между формантами по частоте больше, чем у мужских, так как у них меньше длина акустической полости (см. аналогию по формуле 1.2). На спектре сигнала форманты представляют собой максимумы в окрестностях некоторых частот.

На основе значений формантных частот вокализованных звуков часто реализуются и распознаватели вокализованных звуков, и синтезаторы искусственной речи. Для описания основных свойств гласного звука достаточно определить частоты первой и второй формант (*F1* и *F2*) [52].

Оценка формантных частот связана с рядом трудностей: в слитной речи у мужских голосов между *F2* и *F3* часто появляется дополнительный спектральный максимум; у женских и детских голосов нередко отсутствует *F2* [63].

Одним из параметров спектра фрагмента РС, используемым при решении задачи сегментации, является частота центра тяжести спектра (Spectral Center of Gravity, SCG). В частности, в [64] данный параметр применяется в качестве вспомогательного для разделения альвеолярных фрикативных ([s], [z]) и палатальных [sh], [zh] согласных.

В итоге, речевой сигнал может быть описан большим количеством временных, спектральных, энергетических параметров. Вопрос выбора набора параметров для построения алгоритма сегментации является нетривиальным и представляет большую сложность и важность.

1.2 Анализ основных методов решения задачи сегментации речевого сигнала

1.2.1 Спектральный анализ речевого сигнала

При спектральном анализе РС в общем случае сигнал подвергается дискретному преобразованию Фурье, затем для полученного спектра производится логарифмическое изменение масштаба в пространствах амплитуд и частот (мел-преобразование частоты), выполняется сглаживание спектра и выделяется его огибающая. Это позволяет учитывать понижение информативности высокочастотных составляющих сигнала, а также логарифмическую чувствительность человеческого слуха.

Так, в работе [13] сегментация речевого сигнала на фрагменты, соответствующие фонемам (без их классификации), осуществляется на основе анализа сонограммы, преобразованной механизмом латерального торможения и с учетом математической модели восприятия речи человеком (мел-преобразование частоты, локальное взвешенное интегрирование в частотной области). Границы сегментов определяются по факту значительного изменения спектральной структуры сигнала (рисунок 1.7): считая, что момент начала текущего сегмента

известен (это может быть момент начала РС), концом сегмента считается момент, в который мера близости среднего спектра и каждого профиля превышает некоторый порог.

В каждый момент времени текущего сегмента средний спектр вычисляется рекуррентно [13]:

$$\bar{S}_i(\omega) = \frac{\bar{S}_{i-1}(\omega)\cdot(i-1) + S(\omega, \mathrm{t}_i)}{i}, \quad \bar{S}_0(\omega) = S(\omega, \mathrm{t}_0), \tag{1.3}$$

где $\bar{S}_i(\omega)$ – средний амплитудный спектр на интервале $[t_0, t_i]$, $S(\omega, t_i)$ – профиль сонограммы в момент времени t_i.

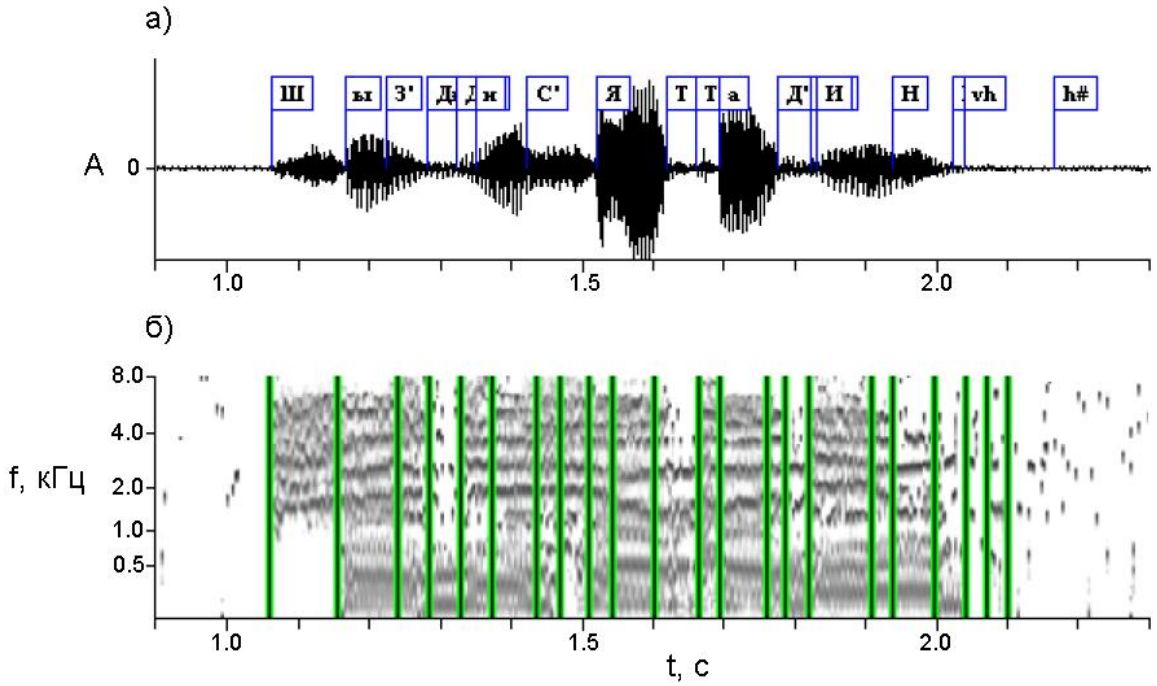


Рисунок 1.7 – Пример автоматической сегментации РС (фраза «Шестьдесят один») на основе анализа сонограммы [13]: а) осцилограмма с ручной разметкой; б) нормированная сонограмма с автоматически найденными границами сегментов

В работе [65] для классификации фонем авторы также используют спектральные характеристики. Спектр Фурье вычисляется на окнах длительностью T = 0,05 с и затем разбивается на 19 равных по частоте интервалов: длительность каждого интервала 500 Гц, общий диапазон от 0 до 5000 Гц (т. к. применяется частота дискретизации 10 кГц), коэффициент

перекрытия интервалов 0,5. Далее в каждом спектральном интервале вычисляются средняя частота ω_i и средняя интенсивность P_i, которые в дальнейшем используются как параметры для распознавания фонем:

$$\omega_i = \frac{\sum_{j=G_i^{(0)}}^{G_i^{(1)}} jS_j}{T\sum_{j=G_i^{(0)}}^{G_i^{(1)}} S_j}; \quad P_i = \frac{1}{G_i^{(1)} - G_i^{(0)}} \sum_{j=G_i^{(0)}}^{G_i^{(1)}} S_j, \tag{1.4}$$

где $G_i^{(0)}$, $G_i^{(1)}$ – соответственно левая и правая границы частотного интервала, S_j – интенсивность j-й гармоники спектра. Описанная методика в [65] применена для классификации гласных и звонких согласных фонем в потоке слитной речи: вероятность правильного распознавания фонем составила 83%.

В работе [64] для разделения альвеолярных фрикативных ([s], [z]) и палатальных ([sh], [zh]) также рассматриваются спектры звуков: палатальные характеризуются компактным спектром, основной пик которого лежит на относительно низкой частоте; у альвеолярных же звуков пик находится на более высокой частоте. Использование этого свойства позволило получить точность разделения этих звуков на уровне 98,5%. В этой же работе для выделения губных и губно-зубных фрикативных согласных предлагается использовать свойство относительной равномерности и небольших амплитудных значений их спектров.

Для автоматической классификации сегментов речи на вокализованные и невокализованные группы звуков в [66] наряду с использованием особенностей корреляционной функции вокализованных фрагментов РС рассматривается распределение энергии различных звуков по частотному спектру. Для характеристики распределения энергии сегмента РС между низкочастотными (до 2 кГц) и высокочастотными (более 2 кГц) поддиапазонами используется относительный полосный энергетический критерий. Так как частота ОТ лежит в пределах от 50 до 500 Гц, а вторая формантная частота не превышает 2 кГц, то в низкочастотной части спектра у вокализованных сегментов имеется большее сосредоточение энергии, нежели у невокализованных. Критерием сегментации

при этом выступает значение отношения сумм квадратов амплитуд последовательности отсчетов спектра РС для обоих поддиапазонов.

Спектральное оценивание

Задача различения фрагментов речевого сигнала в предположении его авторегрессионной природы, может быть решена на основе оценок спектра с помощью критерия минимума информационного рассогласования (МИР) в метрике Кульбака-Лейблера, из которого определяется асимптотически оптимальная решающая статистика для различения сигналов [41]:

$$\gamma_{x,r} \triangleq \frac{1}{F}\sum_{f=1}^{F}\left(\frac{G_x(f)}{G_r(f)} + \ln\frac{G_r(f)}{G_x(f)}\right) \to \min\Big|_{r=\overline{1,R}}, \qquad (1.5)$$

где $G_x(f)$ – выборочная оценка СПМ распознаваемого сигнала x; $G_r(f)$ – выборочная оценка СПМ r из словаря; F – половина частоты дискретизации; R – размер словаря. Для каждой реализации распознаваемого сигнала вычисляется R значений решающей статистики (1.5). Решение принимается по критерию минимума решающей статистики.

В работе [41] данный подход применен для распознавания слов из словаря. При этом слова подвергаются сегментации на участки равной длительности, поэтому решающая статистика для целого слова вычисляется из значений решающей статистики, полученных для каждого отдельного сегмента слова:

$$\gamma_{x,r} = \frac{1}{L}\sum_{i=1}^{L}\gamma_{x,r}^{(i)}, \qquad (1.6)$$

где $\gamma_{x,r}^{(i)}$ – информационное рассогласование между сегментом i сигнала x и сегментом i сигнала r из словаря; L – количество сегментов. В результате исследования разработанного алгоритма для словаря из десяти слов, соответствующих цифрам от нуля до девяти, вероятность дикторонезависимого распознавания составила 0,8.

В работе [67] для сегментации произвольного РС на аллофоны предлагается функция однородности, основанная на среднем значении логарифмического спектра мощности. Проблемы для работы алгоритма, по утверждению его авторов, составляют длинные шипящие, на которых выявляются ложные границы. Кроме того, к сегментируемым записям выставляется условие чередования гласных и звонких согласных, выполняемое далеко не для всех слов.

Говоря о методах, используемых в спектральной области, необходимо еще раз подчеркнуть отрицательное влияние вариативности произнесения речи. Под влиянием различных факторов – эмоционального настроя, физического состояния – даже для одних и тех же фраз или слов человек производит РС, значительно различающиеся по спектрально-временным характеристикам. Помимо этого, в результате изменения темпа и громкости произнесения изменяется также взаимопроникновение соседних звуков [68]. Ввиду сказанного необходимым является использование синхронизируемых со структурой РС алгоритмов сегментации, при этом должна быть обеспечена устойчивость алгоритмов к вариативности произнесения речи.

1.2.2 Кепстральный анализ речевого сигнала

Большинство современных автоматических систем распознавания речи основано на извлечении характеристик текущего состояния речевого тракта человека, а не сигнала возбуждения, так как получаемые в первом случае параметры РС лучше выполняют дистинктивную функцию. Для отделения сигнала возбуждения от сигнала речевого тракта прибегают к кепстральному анализу [34].

Кепстральный анализ основан на гомоморфной обработке речи и позволяет выделить в РС генераторную и фильтровую части [50]:

$$C_S(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \ln\left(S(\omega)\right)^2 e^{i\omega q} d\omega, \qquad (1.7)$$

где $S(\omega)$ – амплитудный спектр речевого сигнала.

Существует несколько методов кепстрального анализа речевых сигналов, применяемых для выделения из них векторов параметров. В первую очередь, к этим методам относятся:

- метод кепстральных коэффициентов линейного предсказания (LPCC – Linear Prediction Cepstral Coefficients) [69, 70];
- метод коэффициентов перцептивного линейного предсказания (PLP – Perceptual Linear Prediction) и робастный метод (PLP-RASTA) [71];
- метод мел-частотных кепстральных коэффициентов (MFCC – Mel Frequency Cepstral Coefficients) [34].

К примеру, в широко распространенном расширении среды MATLAB – пакете функций VoiceBox, специализированных для работы со звуком, – основной акцент делается на использовании коэффициентов MFCC [72].

В уже упоминавшейся работе [14] коэффициенты MFCC используются для реализации сегментации речевого сигнала на ШФК (см. выше рисунок 1.2).

Коэффициенты MFCC не отражают изменения сигнала, поэтому для выявления динамических особенностей используются производные кепстров, образующие дельта-кепстральные коэффициенты DCC (Delta-Cepstral Coefficients) и дельта-дельта-кепстральные коэффициенты DDCC (Delta-Delta-Cepstral Coefficients , «коэффициенты ускорения») [73].

Примером использования кепстрального анализа в задаче фонемной сегментации речевого сигнала является работа [74]. В ней используется 20 мел-частотных треугольных фильтров и вычисляются кортежи по 12 коэффициентов MFCC. Непосредственно для сегментации применяется байесовский информационный критерий. Алгоритм был протестирован на 10 предложениях русского языка, произнесенных мужским голосом. Результаты тестирования приведены в таблице 1.1. Как отмечено автором, для более точного определения границ фонем необходимы дополнительные процедуры, основанные на анализе результатов сегментации и направленные на классификацию сегментов на гласные, смычные согласные, щелевые согласные и т. д.

Таблица 1.1. Результаты автоматической сегментации, основанной на байесовском информационном критерии, с MFCC-параметризацией РС [74]

Верно определенные границы переходов	Пропущенные границы	Неверно определенные границы
414	94	101

Гребенка из 40 треугольных фильтров схожим образом используется для вычисления вектора из 12 MFCC-коэффициентов и в работе [75]. Каждый временной интервал анализа имеет длительность 20 мс с перекрытием 10 мс.

В работе [76] предложен метод фонемной сегментации, основанный на мере спектрального перехода (Spectral Transition Measure, STM):

$$STM(m) = \frac{\sum_{i=1}^{D} a_i^2(m)}{D},$$
$$a_i(m) = \frac{\sum_{n=-I}^{I} c_i(n+m) \cdot n}{\sum_{n=-I}^{I} n^2}, \quad (1.8)$$

где m – порядковый номер текущего окна; D – размерность вектора признаков (количество используемых MFCC-коэффициентов); $c_i(m)$ – i-ый коэффициент MFCC, вычисленный для окна m; $a_i(m)$ – коэффициенты регрессии (скорости изменения вектора коэффициентов MFCC); I – количество окон, используемых для вычисления коэффициентов регрессии $a_i(m)$ (считая раздельно для окон слева и справа от текущего). В описываемой работе авторами применены следующие параметры алгоритма: количество используемых MFCC-коэффициентов D=10; количество используемых смежных окон с каждой стороны от текущего I=2; длительности окон 40 мс; шаг окон 10 мс. В таблице 1.2 приведены данные, полученные при обработке алгоритмом сегментации базы из 4620 предложений, произнесенных 462 дикторами.

Таблица 1.2. Результаты автоматической сегментации на основе меры спектрального перехода [76]

	Всего границ (ручная сегм.)	Определенные верно	Пропущенные границы	Ложные границы
Количество	172 460	145 950	26 510	48 566
Процент	100%	84,6%	15,4%	28,2%

Недостатком данного алгоритма является высокая вероятность пропуска границ между гласными и вокализованными согласными в случаях выраженной коартикуляции или плавного звучания речи [77]. В свою очередь, в работе [77] предлагается модифицировать рассмотренный алгоритм: экспериментально установлено, что в случаях указанных пропусков кепстральная картина приобретает вид, близкий к линейной зависимости (рисунок 1.8). Поэтому вводится мера кепстральной гладкости:

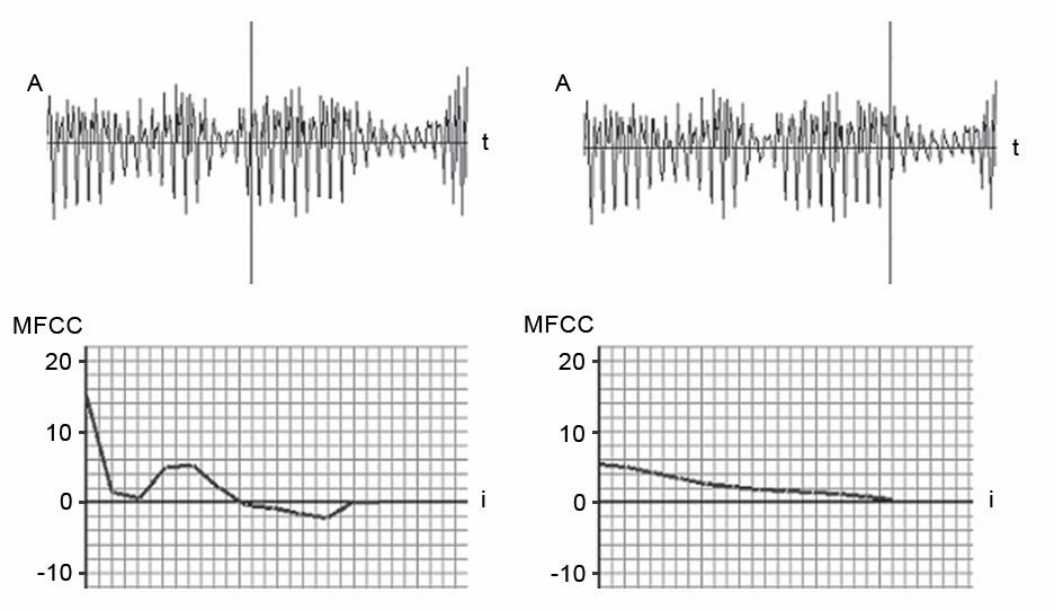
$$CSM(m) = \begin{cases} \dfrac{c_{\max}(m) - c_{\min}(m)}{\sum_{i=1}^{D-1} \left| c_{i+1}(m) - c_i(m) \right|} & , \; c_{\max}(m) \neq c_{\min}(m) \\ 1 & , \; c_{\max}(m) = c_{\min}(m) \end{cases} \tag{1.9}$$

где m – номер текущего окна; $c_i(m)$ – i-ый коэффициент MFCC, вычисленный для окна m; $c_{max}(m)$ – максимальный коэффициент MFCC для окна m; $c_{min}(m)$ – минимальный коэффициент MFCC для окна m; D – размерность кепстра.

Результаты сегментации данным доработанным алгоритмом STM + CSM в сравнении с исходным алгоритмом STM представлены в таблице 1.3.

Таблица 1.3. Результаты автоматической сегментации алгоритмами STM и STM+CSM [77]

	Определенные верно границы	Пропущенные границы	Ложные границы
STM	82,9%	17,1%	20,2%
STM+CSM	90,9%	9,1%	22,3%



а *б*

Рисунок 1.8. Пример коэффициентов MFCC: а) на границе между согласными звуками; б) на границе между согласным и гласным звуком [77]

По приведенной таблице 1.3 видно, что и усовершенствованным алгоритмом сегментации допускается значительное число пропусков и ложных границ, так как в ряде случаев при переходе от фонемы к фонеме не происходит значимых изменений ни по мере спектрального перехода, ни по мере кепстральной гладкости.

1.2.3 Применение вейвлет-преобразования в обработке речевых сигналов

Преобразование Фурье и параметризация коэффициентами линейного предсказания не приспособлены для анализа нестационарных сигналов, так как в данном случае теряется информация о временных особенностях сигнала [53].

В РС есть как фонемные фрагменты с относительно медленным изменением спектрального представления, так и участки быстрой перестройки речевого аппарата (межфонемные переходы, взрывные фонемы) и, соответственно, быстрого изменения спектра сигнала. Такие фрагменты нестационарности делают обоснованным применение вейвлет-анализа для изучения свойств речевого сигнала. Получаемые в результате вейвлет-преобразования вейвлет-спектрограммы содержат информацию и о формантных частотах, и о

гармонической структуре исходного речевого сигнала [78], так как базисные функции вейвлет-разложения обладают способностью выявлять в анализируемом сигнале как частотные, так и временные характеристики, что в результате позволяет выделять и локализовать временные особенности РС [53].

Переход от одних фонем к другим обуславливается изменением конфигурации речевого аппарата, что отражается в резком изменении вейвлет-коэффициентов на одном или нескольких масштабах разложения РС [79].

Вейвлет-разложение РС представляет собой сумму [16]:

$$\begin{aligned} f(t) &= \sum_{k=0}^{N/2^n-1} s_{nk}\varphi_{nk} + \sum_{j=1}^{N/2^n-1} d_{jk}\psi_{jk}, \\ \varphi_{nk} &= 2^{n/2}\varphi(2^n t - k), \text{ где } n,k \in \mathbb{Z}, \\ \psi_{jk} &= 2^{j/2}\psi(2^j t - k), \text{ где } j,k \in \mathbb{Z}, \end{aligned} \tag{1.10}$$

где N – длина анализируемого фрагмента РС; n – число уровней разложения; s_{nk}, a_{jk} – коэффициенты аппроксимации и детализации вейвлет-разложения; φ – масштабная функция; ψ – базисный вейвлет, t – время.

Существенным является выбор базисного вейвлета: он должен позволять описывать стационарные фрагменты РС сравнительно малым числом ненулевых коэффициентов. В алгоритмах сегментации допустимо использование сразу нескольких базисов с последующим объединением результатов, полученных по каждому из них [80]. В качестве оконных функций, применяемых как модулирующих для вейвлетов, в анализе речевых сигналов широкое применение получили окна Гаусса (вейвлет Морлет [81]), Хемминга и Хеннинга [78]. Вейвлет MHat («Мексиканская шляпа») популярен в задачах анализа образов, но неприменим в речевом анализе из-за низкого разрешения формантных частот. В упоминавшейся выше работе [75] для разложения используется вейвлет Морлет:

$$\psi(t) = \exp(-i\omega_0 t)\exp(-\frac{t^2}{2}) \tag{1.11}$$

В работе [33] сегментация РС, осуществляемая на основе дискретного вейвлет-преобразования, показывает надежность 85% независимо от типов фонем. В данном случае сегментация основана на анализе ошибки аппроксимации сигнала: производится декомпозиция сигнала с помощью алгоритма Маллата с последующим восстановлением сигнала. Для 10-уровневого разложения используются вейвлеты Симлета. Ошибка аппроксимации вычисляется как:

$$e(t) = \left\| S(t) - \tilde{S}(t) \right\|, \tag{1.12}$$

где $S(t)$ – исходный сигнал, $\tilde{S}(t)$ – восстановленный сигнал. Далее сигнал ошибки подвергается фильтрации ФНЧ с частотой среза $0{,}025F_s$, пропорциональной частоте дискретизации F_s сигнала. Результат фильтрации сигнала $e(t)$ для слова «Авангард» представлен на рисунке 1.9.

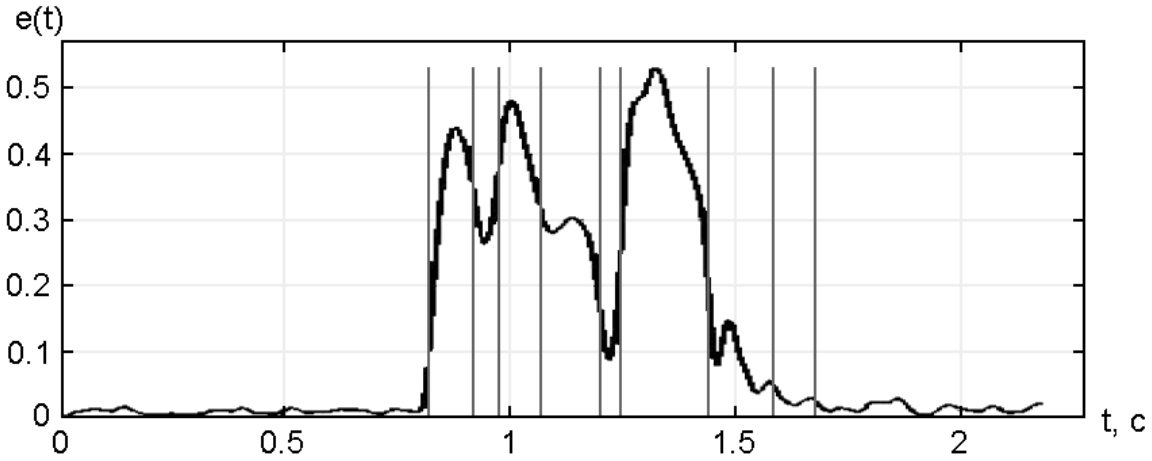


Рисунок 1.9. Функция ошибки аппроксимации и метки ручной сегментации [33]

При автоматической сегментации принятие решения о моментах сегментации строится по скорости изменения функции ошибки $e(t)$. Для этого вычисляется функция скорости изменения ошибки с оптимальной квадратичной интерполяцией и производится поиск максимумов получаемой функции.

При испытаниях алгоритма количество ложных моментов сегментации составило 9% от количества межфонемных переходов согласно эталонной ручной сегментации, а количество пропущенных переходов – 6% от той же величины.

Аналогично, в работе [16] сигнал разбивается на перекрывающиеся участки, к каждому из которых применяется дискретное вейвлет-преобразование. Для каждого *i*-го участка и уровня декомпозиции *n* определяется энергия:

$$E_n(i) = \sum_{i=1}^{2^n-1} d^2_{n,j+2^{n-1}i}, \text{ где } i = 0,...,2^{-M} N - 1 \qquad (1.13)$$

где d_{jk} – коэффициенты детализации вейвлет-разложения; $j = \overline{1, N}$; $k = \overline{0, N / 2^n - 1}$; N – количество отсчетов фрагмента РС; M – количество уровней декомпозиции.

Далее производится сглаживание $E_n(i)$ и, наконец, для определения скорости изменения энергии вычисляется производная $R_n(i)$. Результаты экспериментов, проведенных в [16], показывают незначительную разницу в эффективности использования вейвлетов Майера, Добеши 16, Добеши 8, Симлета 6 порядка – делается вывод о возможности применения перечисленных вейвлетов в качестве базиса разложения с возможным будущим объединением результатов для повышения уровня распознавания границ сегментов.

Важно обратить внимание, что рассмотренные методы вейвлет-сегментации не предусматривают возможности типизации выделяемых сегментов, т. е. их соотнесения с основными типами звуков.

1.2.4 Корреляционный анализ речевого сигнала

Корреляционный анализ сигналов находит широкое применение в вопросах сегментации РС, так как позволяет оценить величину энергии анализируемого фрагмента, факт наличия вокализации, ее частотную локализацию [66]. Главный максимум автокорреляционной функции (АКФ) соответствует энергии анализируемого фрагмента РС, скорость его убывания характеризует наличие шумовой составляющей в РС. Для вокализованных фрагментов АКФ имеет характерный вид, схожий с видом АКФ прямоугольного радиоимпульса.

По значению первого локального максимума АКФ можно судить о наличии вокализации и о среднем значении длительности периода ОТ в рассматриваемом речевом фрагменте [8, 82]. Автокорреляционным методом осуществлялось

извлечение частоты основного тона при создании фонетической базы данных Института русского языка РАН [83].

АКФ относительно устойчива к шумам, однако, для оценивания периодичности вокализованного фрагмента требует достаточно большого интервала анализа. В результате, при рассмотрении переходных участков вокализованных сегментов, сопровождаемых быстрым изменением структуры и частоты колебаний голосовых связок, локальные максимумы становятся невыраженными. Для преодоления данного недостатка в алгоритмах [84, 85] при определении перечня возможных значений периода ОТ используется нормированная кросс-корреляционная функция (НККФ). В предположении, что интервал дискретизации РС $T=1/F_s$ и длительность интервала анализа t, вводится переменная $z=t/T$, тогда НККФ вычисляется как:

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}}, \quad k = \overline{0, K-1}, \quad m = iz, \quad i = \overline{0, M-1}$$
$$e_j = \sum_{l=j}^{j+n-1} s_l^2, \tag{1.14}$$

где i – номер интервала анализа, k – задержка, n – количество отсчетов РС на интервале анализа. Длительность интервала анализа выбирается приблизительно равной ожидаемому значению периода ОТ.

В силу явных различий общего вида АКФ для периодических и случайных процессов, корреляционный анализ также применяется для классификации речевой активности на «тон/не тон» (наличие или отсутствие вокализации). На рисунке 1.10 из [66] показаны примеры АКФ для периодического и шумового сигналов длительностью 30 мс и частотой дискретизации 8 кГц. Заштрихованная область обозначает интервал, на котором не производится поиск значения основного тона: интервал вводится для борьбы с вероятными ложными максимумами, возникающими из-за нестационарности сигнала.

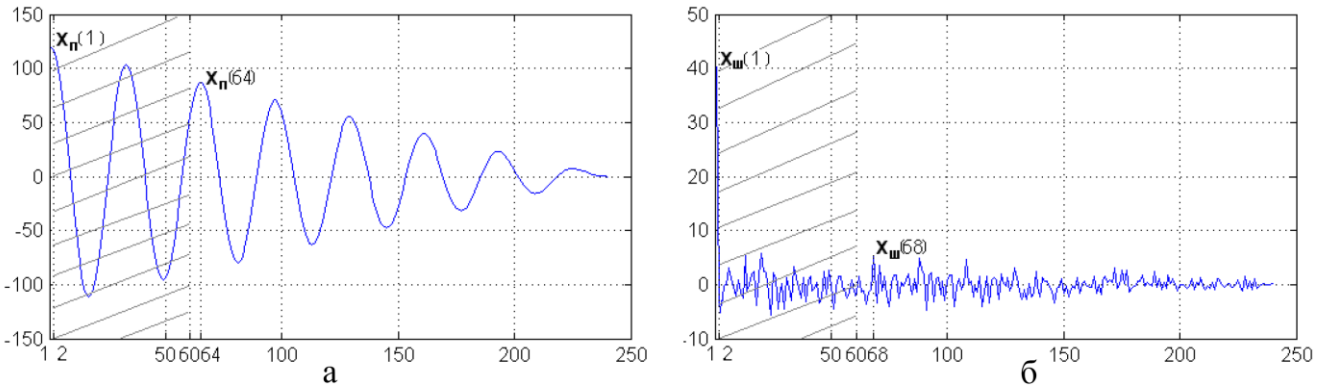


Рисунок 1.10 – АКФ: а) периодического $x_п(n)$; б) шумового сигнала $x_ш(n)$ [66]

Корреляционные методы анализа показывают хорошие результаты при рассмотрении стационарных вокализованных фрагментов квазигармонической структуры. Однако работа алгоритмов значительно усложняется наличием звуков со сложной структурой колебаний ОТ. Необходимость достаточно длительного интервала анализа приводит к ошибкам работы алгоритмов на фрагментах быстрой перестройки артикуляционного аппарата. Кроме того, корреляционные методы анализа требуют значительных вычислительных затрат [86], поэтому усилия исследователей направлены также на повышение быстродействия корреляционных алгоритмов обработки РС [87].

1.3 Базовые задачи сегментации речевых сигналов

Практически во всех приложениях обработки речевых фонограмм можно выделить некие предварительные операции, результатом которых в целом является оценка основных характеристик фонограммы, скорости их изменения во времени. Такие операции реализуются технологическими алгоритмами и могут быть использованы в различных речевых приложениях. Результаты работы технологических алгоритмов в дальнейшем, в зависимости от поставленной задачи, используются функциональными алгоритмами, представляющими определенное конечное речевое приложение (рисунок 1.11).

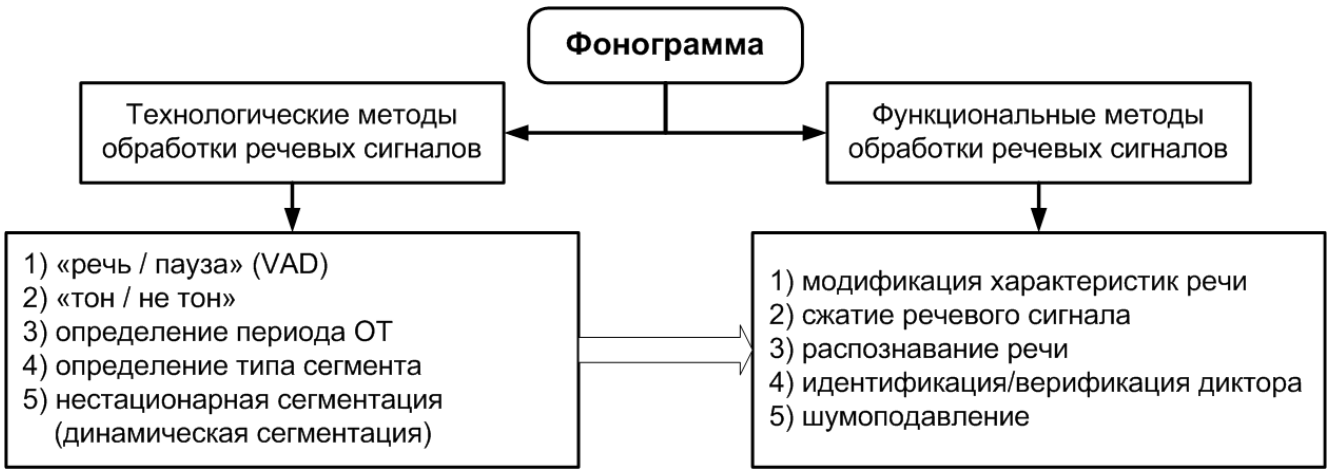


Рисунок 1.11 – Технологические и функциональные алгоритмы обработки РС

1.3.1 Определение границ речевой активности

Речь является способом обмена информацией между людьми, производимая речевым аппаратом и передаваемая в естественных условиях посредством звуковых волн. В голосовой связи временные интервалы, на которых сигнал содержит речевую информацию, относят к участкам речевой активности. И наоборот, временные интервалы, не содержащие речевой информации, вне зависимости от наличия или отсутствия фоновых шумов, относят к участкам пауз.

Алгоритм определения временных границ речевой активности и в зарубежной, и в отечественной литературе обозначают сокращением VAD (Voice Activity Detection). VAD-алгоритм позволяет произвести сегментацию сигналов на два типа сегментов: активности *A* (activity) и пауз *S* (silence) (рисунок 1.12).

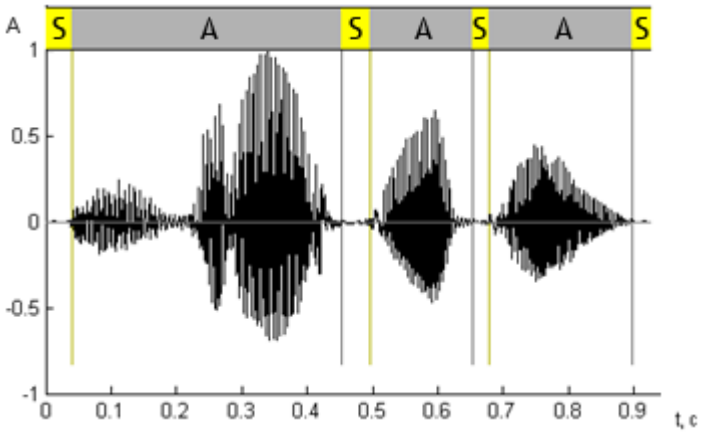


Рисунок 1.12 – Сегментация VAD

Одним из наглядных примеров применения VAD-алгоритма является сотовая система подвижной радиосвязи стандарта GSM (Global System for Mobile Communications, глобальная система мобильной связи) [88]. Обработка речи осуществляется в соответствии с принципом прерывистой передачи DTX (Discontinuous Transmission), что позволяет включать передатчик только на время активной речи пользователя и отключать его в паузах и в конце разговора. DTX управляется VAD-детектором, необходимым для обнаружения и выделения интервалов наличия и отсутствия речи даже при низких ОСШ около 0 дБ. В GSM-системах VAD-детектор играет ключевую роль в снижении потребления, так как при монологе средняя активность речи говорящего ниже 50%, а при диалоге активность участника может снижаться до 30% от времени разговора [89].

Алгоритм VAD GSM достаточно ресурсоемкий с точки зрения машинного времени, менее требовательным является алгоритм Рабинера-Самбурома [8], в котором для формирования порога в качестве основных параметров используются функция числа переходов через ноль Z_n и функция среднего значения M_n РС $x(m)$, вычисленные с использованием окна длительностью 10 мс (N отсчетов), формулы (1.15) и (1.16). Для оценки статистических параметров фонового шума алгоритм рассматривает первые 100 мс РС (считается, что на этом интервале речевая активность отсутствует).

$$Z_n = \sum_{m=-\infty}^{\infty} \left| \operatorname{sgn}\left[x(m)\right] - \operatorname{sgn}\left[x(m-1)\right] w(n-m) \right|$$

$$\operatorname{sgn}\left[x(n)\right] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \qquad (1.15)$$

$$w(\mathrm{n}) = \begin{cases} 1/2N, & 0 \leq n \leq N-1 \\ 0, & \text{остальные случаи} \end{cases}$$

$$M_n = \sum_{m=-\infty}^{\infty} \left| x(m) \right| w(n-m) \qquad (1.16)$$

Затем с учетом этих характеристик вычисляются соответствующие пороги определения речевой активности. Выделяется сегмент РС, для которого значение M_n превышает верхний порог *ITU* (см. рисунок 1.13). Предполагается, что начало и конец слова лежат вне этого фрагмента. Затем, двигаясь по оси времени в направлении от момента, где M_n впервые превысила порог *ITU,* определяют момент, в котором M_n впервые оказалась меньше нижнего порога *ITL* (точка N_1 на рисунке 1.13). Этот момент выбирается в качестве предполагаемого начала речевой активности. Сходным образом определяется и момент предполагаемого окончания активности N_2.

Данный двухпороговый алгоритм гарантирует, что кратковременные провалы в динамике среднего значения РС не приведут к ложному выделению моментов начала и конца слова. Задача первого шага – получить данные о том, что начало и конец слова расположены вне интервала от N_1 до N_2. Следующий шаг состоит в перемещении влево от N_1 (вправо от N_2) не более чем на 25 интервалов анализа и сравнении числа переходов через нуль с порогом *IZCT* (см. рисунок 1.13). Если порог *IZCT* превышается в 3 или более раз, то метка N_1 начала слова переносится в момент первого превышения порога (аналогичным образом корректируется оценка N_2 момента окончания речевой активности).

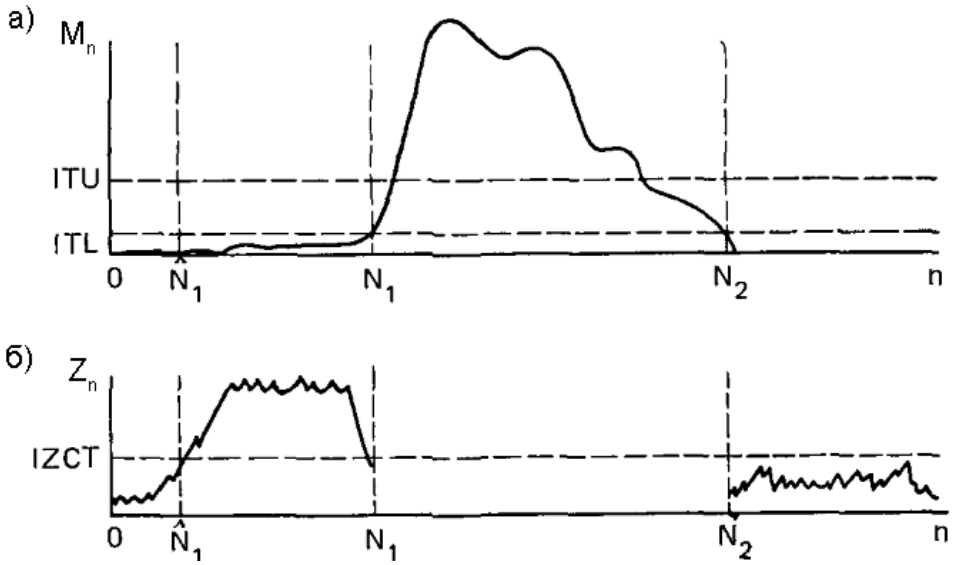


Рисунок 1.13. Сегментация VAD-алгоритмом: а) функция среднего значения РС; б) функция среднего значения пересечения нуля [8]

Достоинствами алгоритма является простота и достаточная точность определения границ речевой активности даже на звуках с малой энергией.

Существует стандартизованный алгоритм VAD в составе кодека речи, известного под названием ITU Standard – G.729 [90]. Для определения границ активности речи VAD-алгоритм G.729 использует четыре характеристики фонограммы: полнодиапазонную энергию, энергию низких частот, частоту пересечений нуля и коэффициенты спектра сигнала. Вычисляются различия между каждой из этих характеристик и их средние значения. Эти средние значения обновляются на участках пауз. Решение о наличии или отсутствии речи выносится по многомерному вектору параметров.

Алгоритм G.729 демонстрирует лучшую эффективность [91], нежели VAD, применяемый в Half Rate GSM (скорость кодирования не выше 4 кбит/с). Более того, в работе [92] предлагается усовершенствование алгоритма G.729: для выделения огибающей спектра вместо стандартного алгоритма линейного предсказания предлагается использовать алгоритм TE-LPC (True Envelope Linear Predictive Coding). Алгоритм основан на ограниченной по диапазону интерполяции вычисленного спектра сигнала [93]. Разница между алгоритмами LPC и TE-LPC показана на рисунке 1.14.

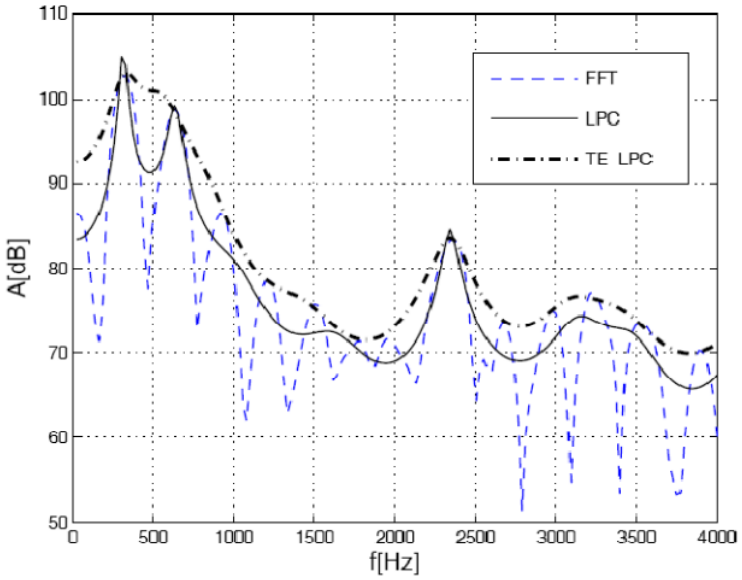


Рисунок 1.14 – Огибающая спектра (FFT): алгоритмы LPC и TE-LPC [93]

В работе [94] предлагается VAD-алгоритм, имеющий раздельные критерии для выделения активной и неактивной речи:

– для активных фрагментов речи для принятия решения используется отношение энергии сигнала, пропущенного через специальный фильтр, к полной энергии сигнала. АЧХ фильтра модулирована функцией Гаусса на частотах гармоник РС;
– для неактивных фрагментов речи для принятия решения используется отношение энергии высокочастотной и низкочастотной составляющих сигнала.

Работа VAD-алгоритмов затрудняется при наличии значительных фоновых шумов, поэтому в таких ситуациях применяются более сложные подходы. Так, в работе [95] предлагается VAD-алгоритм с адаптивным порогом, который показывает хорошие результаты даже при изменяющемся во времени отношении сигнал-шум.

Многие современные разработки алгоритмов выделения речевой активности являются коммерческими и не имеют в открытом доступе достаточного для воспроизведения работы описания. Основными проблемами существующих VAD-алгоритмов является ненадежная работа при повышении уровня фонового шума, а также требуемые значительные вычислительные ресурсы.

1.3.2 Выделение основных типов речевой активности

Необходимость разделения речевой активности на вокализованные/невокализованные/взрывные/шумные звуки связана с кардинально различающимися сигнальными свойствами данных фрагментов РС и, соответственно, необходимостью применения различных методов обработки в функциональных речевых алгоритмах.

Помимо рассмотренных выше в подразделе 1.2 «Анализ основных методов решения задачи сегментации речевого сигнала» методов для выделения определенных типов звуков могут применяться также более специфические

признаки соответствующих сегментов. Такими признаками являются скорость флюктуации сигнала, средняя мощность и длительность.

Одним из классифицирующих параметров является частота переходов через нуль (ZCR, Zero Crossing Rate): невокализованные фрагменты РС флюктуируют значительно быстрее, и поэтому имеют значительно большее значение ZCR, что позволяет применять данный параметр для сегментации «тон/не тон» [76, 77, 40].

С энергетической точки зрения простым, но эффективным параметром для выделения вокализованных звуков является средняя мощность: для вокализованных она обычно принимает сравнительно большие значения [40]. Для более эффективной работы алгоритма сегментации средняя мощность сегментов может рассматриваться в рамках определенных частотных диапазонов [66, 96].

Наконец, для взрывных звуков, произнесение которых связано с резким размыканием препятствия в речевом аппарате, характерна небольшая по сравнению с шумными и вокализованными длительность.

1.3.3 Выделение периодов основного тона

Сегментация РС на периоды ОТ заключается в выделении временных границ каждого отдельного колебания голосовых связок. Частота ОТ непрерывно изменяется в процессе произнесения речи. Можно выделить три первопричины отклонений частоты основного тона от среднего для диктора значения [97]:

- случайные отклонения в длительностях вокальных циклов (джиттер-эффект [98, 99, 100]), отклонения частоты возможны как в большую, так и в меньшую стороны;
- вибрато – периодическое изменение частоты ОТ; характерно для вокальной речи, построенной на частотной модуляции звука; частота модуляции лежит в пределах от 5 до 8 Гц [20, 101];
- плавное увеличение и уменьшение частоты ОТ в процессе речи как инструмент интонационной окраски; обычно с увеличением частоты ОТ, вызываемого увеличением подсвязочного давления, увеличивается также и громкость речи.

Для выделения из речевой фонограммы периодов ОТ существует большое разнообразие алгоритмов, работа которых основана на разных математических подходах [10]. Можно выделить две основные подгруппы алгоритмов оценки периода ОТ [102]:

а. структурные методы оценки периода ОТ, или локальные;

б. алгоритмы без привязки к временной структуре, или интегральные.

Условная классификация алгоритмов оценки периода ОТ РС представлена на рисунке 1.15. Методы, перечисленные в блоках с серым цветом фона, позволяют получить помимо средней («интегральной») оценки величины периода ОТ на некоторой анализируемом интервале РС также текущие значения периода ОТ, привязанные к конкретным отсчетам в РС.

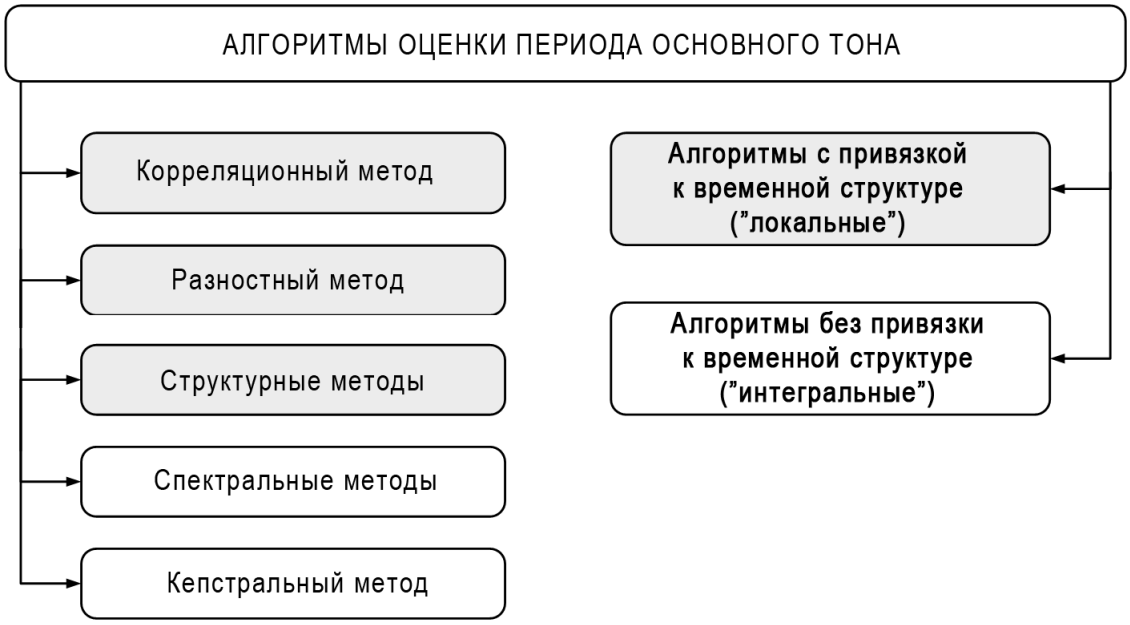


Рисунок 1.15. Классификация алгоритмов оценки периода ОТ речевого сигнала

Суть подгруппы «локальных» алгоритмов заключается в измерении каждого значения периода ОТ на исходной не модифицированной фонограмме. Структурные методы оценки периода ОТ «привязывают» итоговые измерения периодов к некоторым характерным периодически повторяющимся особенностям структуры речевого сигнала, существующим на интервале ОТ-периода (к примеру, к максимальному экстремуму, к точке пересечения через нуль). В результате работы такого алгоритма возможна разметка РС по периодам ОТ, а

также формирование текущей (мгновенной) оценки периода ОТ по фонограмме. Методы данной подгруппы являются эффективным инструментом, используемым в алгоритмах временной сегментации при реализации разметки РС на периоды ОТ.

К группе «интегральных» алгоритмов оценки периода ОТ без привязки к временной структуре относятся следующие методы:

- корреляционный метод,
- «разностный» метод, или метод с использованием кратковременной функции среднего модуля значения разности,
- кепстральный метод,
- спектральные методы,
- пиковый метод (амплитудная селекция) и другие…

Алгоритмы без привязки к временной структуре («интегральные») не являются базовыми алгоритмами сегментации, так как информация о среднем значении периода для задач сегментации, в частности ОТ-сегментации, является дополнительной (второстепенной). В то же время, интегральная оценка частоты ОТ фрагмента РС зачастую применяется в качестве вспомогательного входного значения алгоритмами ОТ-сегментации.

Для выделения текущей частоты ОТ может быть использовано свойство кратности частот гармоник основного тона частоте первой гармоники [103]. На вход алгоритма должен подаваться фрагмент речевого сигнала, длительность которого гарантированно больше длительности периода ОТ. При работе алгоритма производится ряд сжатий спектра в целое число раз. При этом при каждом сжатии частота соответствующей высшей гармоники ОТ будет совпадать с частотой первой гармоники.

Так же выделение частоты ОТ (но не сегментация на отдельные периоды ОТ) может быть осуществлено с помощью узкополосного фильтра, следящего за частотой первой гармоники РС [104]. Ширина полосы фильтра с использованием обратной связи подстраивается под среднюю частоту основного тона диктора.

1.4 Основные выводы по разделу

Рассматривая современные исследовательские направления, касающиеся автоматической обработки РС, можно сделать вывод о важности задачи временной сегментации и повсеместном ее применении. Наиболее распространенными уровнями временной сегментации являются: VAD-сегментация, сегментация РС на фрагменты характерных типов, сегментация на ШФК, фонемная сегментация, сегментация вокализованных фрагментов на периоды ОТ.

При рассмотрении задач сегментации РС речевой аппарат как акустическую систему принято представлять в виде трех функциональных блоков: генератор, вибраторы и резонаторы. Для описания поведения речевого тракта как динамической трубы используют авторегрессионную модель РС.

Среди уровней временной сегментации РС наиболее сложным – и наиболее близким к задаче распознавания речи – является фонемная сегментация. В разделе достаточно детально освещены вопросы воспроизведения русских фонем в процессе речи, в том числе обозначены основные факторы, влияющие на разнообразие вариаций фонем. Приведены примеры существующих классификаций фонем для задач технического рассмотрения речи.

Одним из ключевых вопросов в задаче временной сегментации является выбор подходящего ряда параметров РС, на основе которого будет производиться сегментация. Обзор современных научных источников указывает на огромное многообразие результативно применяемых для сегментации параметров и комбинаций параметров РС.

Базовыми алгоритмами сегментации, имеющими самостоятельное значение и наиболее часто встречающимися, являются алгоритм определения активности говорящего (VAD-алгоритм) и алгоритм выделения отдельных периодов ОТ вокализованных фрагментов речевой активности.

2 ИССЛЕДОВАНИЕ СИГНАЛЬНЫХ ОСОБЕННОСТЕЙ ЗВУКОВ РУССКОЙ РЕЧИ

Наиболее сложной задачей временной сегментации РС является фонемная сегментация. На основе фонемной сегментации можно получить большинство менее детальных уровней сегментации простым объединением сегментов, соответствующих фонемам. Во многих системах автоматического распознавания РС основополагающее значение играет именно сегментация сигнала в соответствии с фонетической транскрипцией [28].

Знание сигнальных характеристик фонем и их групп является неотъемлемым условием разработки алгоритмов многоуровневой временной сегментации. Однако, как было показано выше, существующие классификации и описания фонем русской речи являются несистемными и во многом противоречивыми, что приводит к необходимости дополнительного исследования особенностей русского речевого сигнала.

2.1 Фонетический алфавит: звуки русской речи и их группы

Как средство передачи информации, речь должна описываться конечным числом различимых и взаимоисключающих звуков. При акустической реализации в процессе произнесения человеком звуки могут быть существенным образом видоизменены (см. выше подраздел 1.1.4 «Параметризация сегментов речевого сигнала»). Однако, несмотря на это, при восприятии речи на слух разнообразные реализации звука соотносятся в сознании с одним лингвистическим элементом – фонемой. Отдельный вариант реализации фонемы называют аллофоном.

Аллофонные варианты реализации фонемы разделяют на существенные (такой аллофон не может быть заменен на другой аллофон этой же фонемы) и, соответственно, несущественные. К существенным аллофонам относят:

– основной аллофон фонемы, свойства которого практически не зависят от положения в слове и фонетического окружения: для гласных основной аллофон образуется при отдельном произнесении фонемы; для твердых согласных образуется при произнесении перед ударным [а]; для мягких

согласных образуется при произнесении перед ударным [и];

– комбинаторные аллофоны, обусловленные фонетическим окружением, то есть действием соседних звуков;

– позиционные аллофоны, обусловленные позицией фонемы в слове.

В русском языке, как уже говорилось в первом разделе, выделяют *43* основных звука (фонемы): *6* гласных и *37* согласных. В фонетической транскрипции звуки могут быть записаны с помощью русского фонетического алфавита, основанного на написании букв русского алфавита, либо с помощью символов международного фонетического алфавита. В данной работе используется первый вариант, как наиболее интуитивно понятный, обладающий более компактной записью и наиболее удобный при машинном хранении и обработке данных (в русском фонетическом алфавите не используются специальные диакритические знаки).

Для указания ударного гласного используются символы в верхнем регистре. Из числа некириллических символов в записи звуков используются следующие:

j – для обозначения звука «*й*», например, слово *жёлтый* [жОлтыj];

‘ – одинарная кавычка для обозначения мягкого согласного; фонемы [ч’] [щ’] всегда мягкие, фонемы [ж] [ц] [ш] всегда твердые;

: – двоеточие используется для обозначения долгого звука, например, слово *сжечь* [ж:Эч’].

К основным гласным фонемам относятся:

[а] [э] [и] [ы] [о] [у]

К основным согласным фонемам относятся:

[б] [б'] [в] [в'] [г] [г'] [д] [д'] [ж] [з] [з'] [j] [к] [к'] [л] [л'] [м] [м'] [н] [н'] [п] [п'] [р] [р'] [с] [с'] [т] [т'] [ф] [ф'] [х] [х'] [ц] [ч'] [ш] [щ’] [ж':]

Существование фонемы [ж’:], парной фонеме [щ’], является предметом споров московской и петербургской фонологических школ [51]. В рамках диссертационной работы допускается существование звука [ж’:], встречающегося, например, в словах *вожжи*, *дожди* (при произнесении как «*дожжи*»).

Кроме того, в силу вероятных сигнальных особенностей, в исследовании отдельно рассматриваются дополнительные часто встречаемые реализации основных перечисленных выше звуков (их аллофоны). Во-первых, это длительные согласные звуки [д:] [ж:] [н:]. Во-вторых, аллофоны гласных [105]:

[и^{е}] – фонема [и] в безударном положении не перед мягкими согласными;

[и^{э}] – в предударном слоге звучит на месте гласных *А, О, Э* после мягких согласных, похож на [и] с призвуком [э];

[ы^{э}] – буква *А* в предударном слоге обычно после твердых *Ж, Ш, Ц* перед мягким согласным, буква *Е* в предударном слоге после *Ж, Ш, Ц*.

В ряде исследований учитывается дополнительная подклассификация звуков по их положению относительно соседних звуков. Например, в [28] для ударных гласных: положение между твердыми согласными, между мягким и твердым согласными, между твердым и мягким согласными, между мягкими согласными. Аналогичная подклассификация в указанной статье вводится для безударных гласных; таким образом, количество звуков, подлежащих классификации, возрастает до 77. В диссертационной работе такая подклассификация не производится, так как местоположение звука относительно начала слова, относительно конца слова и относительно ударного гласного в виде отдельных полей хранится в разработанной базе данных. Это позволяет, при необходимости, сделать выборку реализаций звуков с учетом требований по их взаимному расположению с еще более гибкими условиями.

2.2 Основные типы фрагментов речевой активности

В данном подразделе рассмотрены вопросы формирования акустического сигнала различных типов звуков в речевом аппарате человека, отражена сложность состава речи человека, ее изменчивость в зависимости от физиологических особенностей говорящего, его эмоционального состояния. Приводится общая классификация типов сегментов речевой активности в РС.

Естественная человеческая речь имеет сложную структуру и весьма разнообразна. Разнообразие возникает по ряду причин, среди которых:

– различия голосов;

– различия громкости;

– вариации интонации произношения;

– варьирование движения артикуляторов (языка, губ, челюсти, нёба).

Сложность структуры РС является следствием сложности устройства голосового аппарата человека, работа которого так и не была достаточно глубоко исследована из-за трудностей наблюдения за артикуляторами. Кроме того, не стоит забывать, что на форму РС влияет не только физиология строения задействованных в речеобразовании органов, но и форма сигнала, идущего от мозга к мышцам голосового тракта. Самый простой способ исследования свойств голосового тракта – его моделирование набором фильтров. В модели используются источники звука, вызывающие резонанс на так называемых формантных частотах.

При произнесении звуков вибрация голосовых связок является источником возбуждения и вызывает резонанс между голосовыми связками и губами. Так как язык, челюсть, губы, зубы и альвеолярный аппарат (альвеолы – углубления в челюстных костях для помещения зубов; альвеолярный или зубной отросток – та часть челюстной кости, в которой помещаются зубы [106]) двигаются, размеры и места этих резонансов меняются, давая возможность воспроизведения особых параметров звуков. Турбулентный шум в РС возникает в результате прохождения воздуха через сужения речевого тракта. Голосовой аппарат человека в результате действует как линейный фильтр с изменяющимися во времени параметрами.

Исследовав примеры фонограмм речи человека, можно убедиться, что РС состоит из фрагментов нескольких основных типов. Это вокализованные, шумные, взрывные типы, а также паузы между звуками речи. Следует отметить, что достаточно частым явлением является наличие небольшой паузы между частями одного слова – перед взрывным звуком. В таких случаях паузы принято также называть смычками.

Все гласные звуки произносятся при активности только голосовых связок и без создания сужений в голосовом тракте.

При произнесении согласных звуков в речевом аппарате возникают препятствия свободному прохождению воздуха. В результате, согласные, в том числе вокализованные ([б], [н], [р],..), во-первых, имеют меньшую мощность сигнала по сравнению с гласными, а во-вторых, если при произнесении звука воздух с силой проталкивается через значительные сужения, в сигнале появляется значительная шумовая компонента ([з], [с], [ш], [т’]). Кроме того, при произнесении согласного голосовые связки могут быть неактивны – тогда произносится либо чисто шумный звук, возникающий при длительном протягивании воздуха через значительное сужение (ср. вокализованный [з] и шумный [с]), либо так называемый взрывной звук, возникающий при силовом проходе воздуха через резко образуемое отверстие ([п], [т], [к]...).

Таким образом, вокализованные звуки (рисунок 2.1) формируются с участием голосовых связок, шумные глухие (рисунок 2.2) – за счет прохождения воздуха через сужения голосового тракта, а взрывные глухие (рисунок 2.3) – с помощью кратковременного смыкания речевого аппарата, создания в речевых полостях повышенного давления и затем резкого размыкания речевого аппарата. Форма голосового тракта остается неизменной на интервале от 10 до 30 мс. На этом интервале речь можно рассматривать как стационарный случайный процесс. Поэтому большинство алгоритмов предварительной обработки обеспечивают анализ речи на указанном интервале времени.

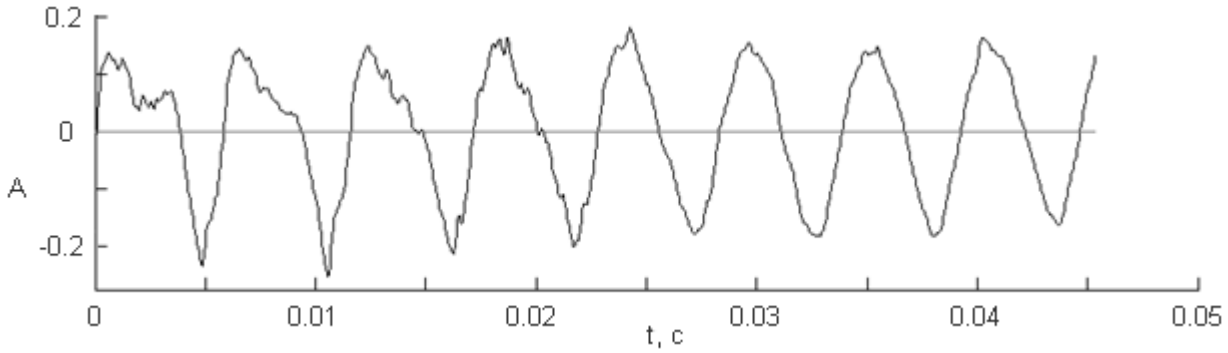


Рисунок 2.1 – Пример вокализованного фрагмента РС, звук [н]

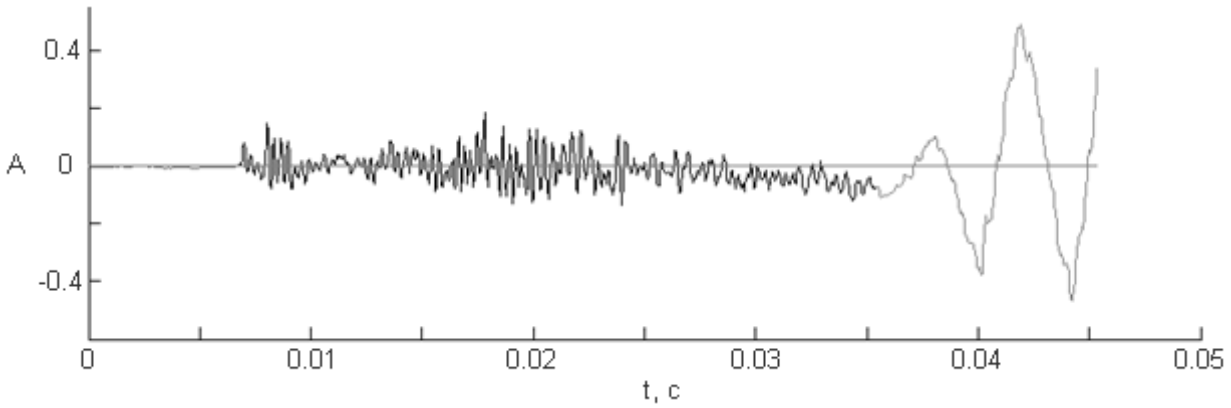


Рисунок 2.2 – Пример шумного фрагмента РС, звук [т’]

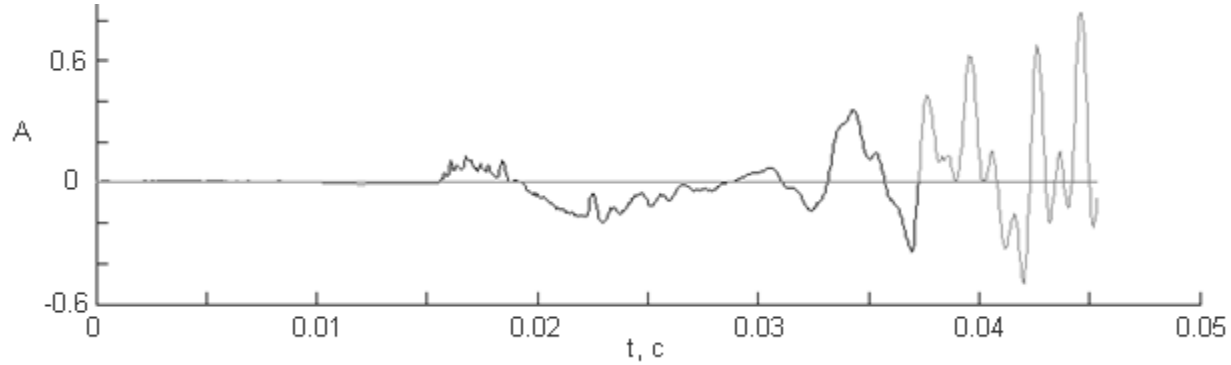


Рисунок 2.3 – Пример фрагмента РС со взрывным звуком: звук [п] между смычкой и ударным [а]

Важно обратить внимание, что вокализованные сегменты РС могут образовывать не только гласные звуки, но и звонкие согласные. Невокализованные сегменты, в свою очередь, представлены глухими согласными, к которым относятся шумные и взрывные звуки. Иерархическая модель структуры РС представлена на рисунке 2.4.

Рисунок 2.4 – Иерархическая модель структуры РС

2.3 Вычисление и анализ ряда сигнальных параметров реализаций фонем русского языка

Для проведения исследования использована специально подготовленная отсегментированная вручную до фонемного уровня база фонограмм русской речи. Детально разработанная методика исследования сигнальных особенностей звуков приведена в Приложении А.

Текущий подраздел начинается с анализа параметров, характерных для всех типов сегментов РС, в то время как последний рассматриваемый параметр – количество переколебаний на периоде ОТ – по определению имеет смысл только для вокализованных звуков. По каждому из параметров описывается процесс его вычисления, приводится описание основных выявленных при рассмотрении параметра закономерностей по группам звуков.

Основные параметры, численно характеризующие текущую звуковую реализацию, и характеристики сигнала, используемые для извлечения параметров, представлены на рисунке 2.5. Некоторые из перечисленных на рисунке параметров рассмотрены далее.

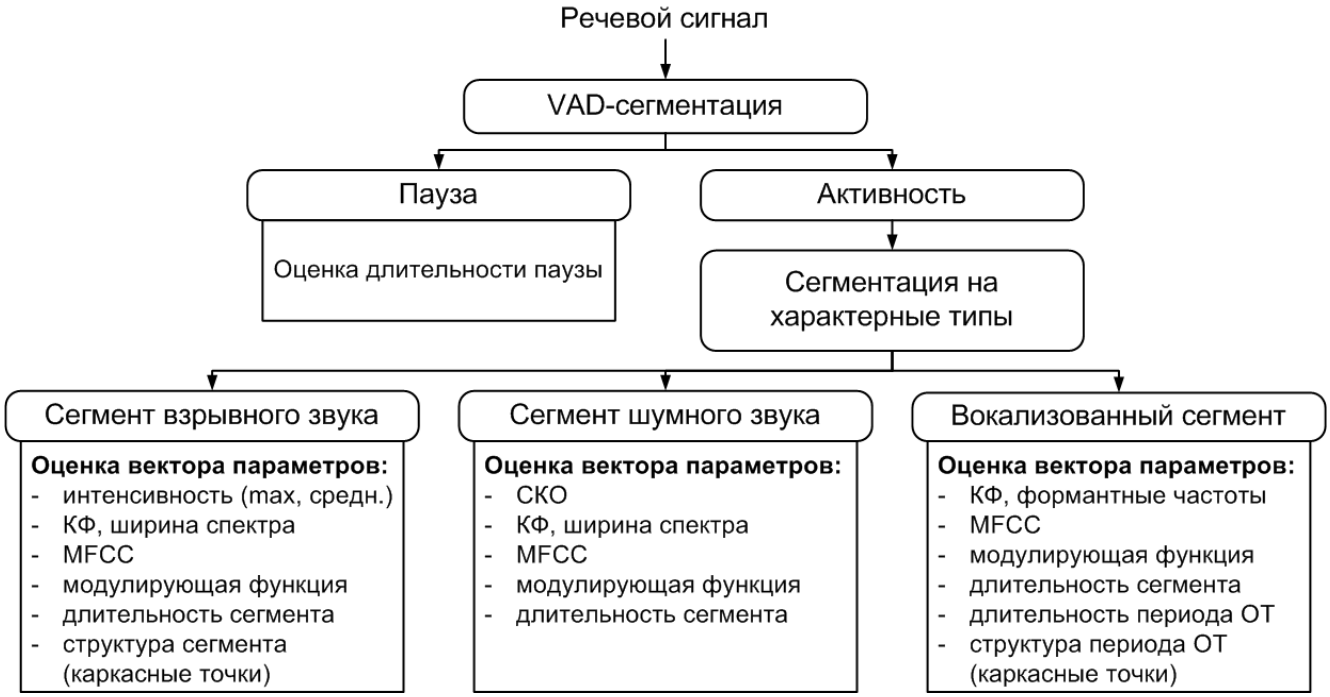


Рисунок 2.5 – Параметры и параметризуемые характеристики основных типов сегментов РС

2.3.1 Длительность звука

Одним из простейших параметров звука является его длительность:

$$\tau = \frac{N}{f_S}, \tag{2.1}$$

где N – количество дискретных отсчетов (сэмплов), которые приходятся на звук, f_S – частота дискретизации PC.

В Приложении Б приведена полная таблица средних длительностей звуков, составленная без учета диктора, ударности гласного и положения звука в слове (таблица Б.1). При таких условиях возможные минимальная и максимальная длительности одинаковых фонем в зависимости от контекста произнесения могут различаться в 2-5 раз. Здесь и далее, если не оговорено специально, длительность вокализованного звука перед паузой или смычкой вычисляется без учета релаксационных колебаний голосовых связок. На рисунке 2.6 средние длительности звуков отображены графически на диаграмме.

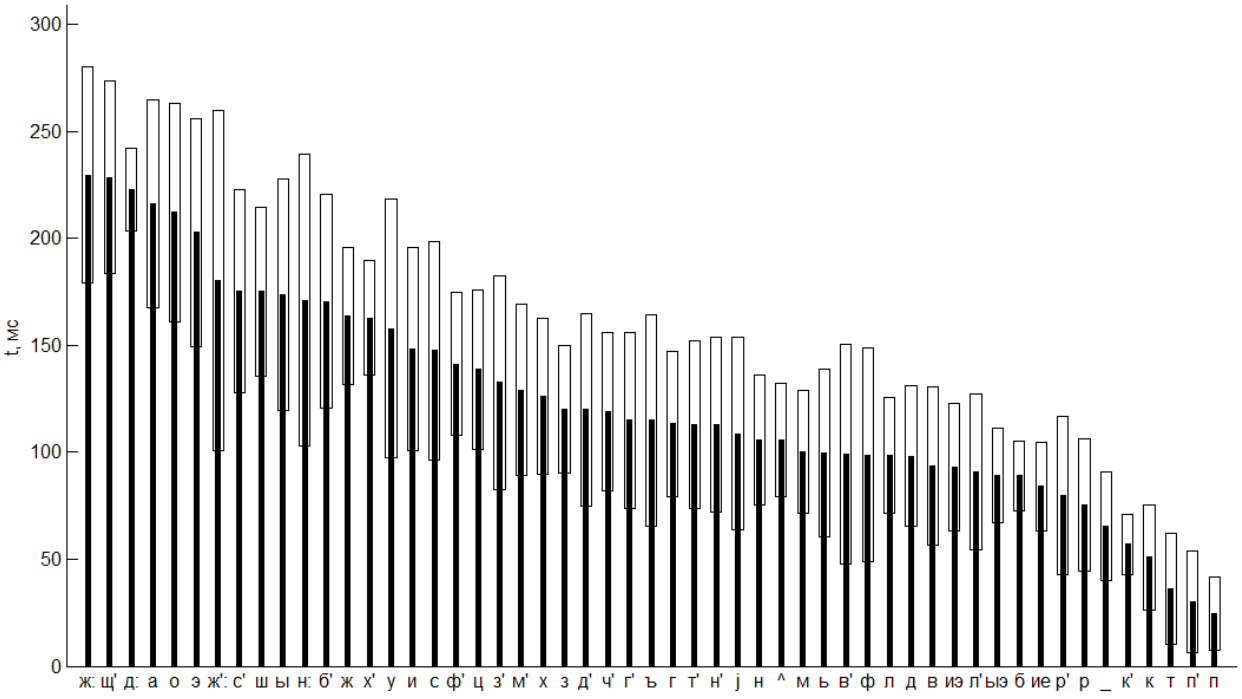


Рисунок 2.6 – Оценки средних длительностей звуков ±σ-размахи

Хранение информации о реализациях фонем в базе данных позволяет делать выборки экземпляров фонем и их параметров по разным условиям. Рассмотрим более детально длительности гласных. Под ударением встречаются только шесть основных гласных звуков, их средние длительности при расположении не в конце слова представлены в таблице 2.1. Значения для гласных в слабых позициях представлены в таблице 2.2. Здесь и далее вычисление стандартного отклонения σ осуществляется на основе несмещенной оценки дисперсии случайной величины:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \qquad (2.2)$$

где x_i – i-й элемент выборки, $\bar{x}$ – среднее арифметическое, N – объем выборки.

Таблица 2.1. Средние длительности ударных гласных не в конце слова

Звук	Средняя длительность, мс	СКО, мс	Кол-во измерений	Мин.	Макс.
[а]	210,9	49,0	96	113,0	332,2
[о]	210,2	51,7	85	107,8	329,3
[э]	204,1	50,3	69	68,0	323,3
[ы]	189,7	42,1	14	136,6	267,9
[у]	187,1	51,5	22	103,6	280,5
[и]	166,5	33,5	38	93,6	249,2

Таблица 2.2. Средние длительности безударных гласных не в конце слова

Звук	Средняя длительность, мс	СКО, мс	Кол-во измерений	Мин.	Макс.
[ы]	108,8	49,8	7	64,7	205,4
[у]	108,2	39,8	25	53,9	244,2
[и]	107,1	29,6	28	53,6	160,5
[^]	105,6	26,6	74	48,4	169,2
[ы^{э}]	89,4	22,1	11	65,2	131,8
[и^{э}]	87,2	20,5	29	46,4	129,1
[ь]	84,3	19,3	27	51,2	120,8
[э]	83,4	0,2	2	83,2	83,5
[ие]	83,2	20,3	38	39,5	121,3
[ъ]	80,9	24,3	58	44,9	159,4

Как видно из приведенных данных, гласные в безударной позиции имеют в среднем примерно в два раза меньшую длительность по сравнению с ударными. Звукам в конце слова характерно возможное наличие затухающих свободных колебаний голосовых связок, имеющих длительность до ~120 мс. В исследовании изначально такие свободные колебания в конце слова или перед взрывным звуком относятся к предваряющему их звуку. Такое решение было принято из-за факта наличия речевой активности и вокализации в этих колебаниях, хотя технически такие фрагменты в силу отсутствия смысловой нагрузки и крайне слабой слышимости можно выделить в отдельный звук сродни смычке или отнести непосредственно к паузе / смычке.

Среди согласных можно выделить группу взрывных звуков [к], [т], [п], наиболее коротких по длительности (рисунок 2.6).

Информация о возможном диапазоне длительности смычки важна при проектировании VAD-алгоритма и позволяет сократить количество ложных фрагментов паузы (см. ниже пункт 3.4.2 «Повышение эффективности энергетического VAD-алгоритм»). Для образования взрывного звука необходимо предварительно создать в речевом тракте значительное избыточное давление, полностью перекрыв выход воздуха на некоторый интервал времени – смычку – длительностью около 30 мс [28]. В свою очередь, согласно результатам измерения длительности смычки в диссертационной работе установлено, что минимальная длительность смычки перед взрывным звуком составляет 20-25 мс, а максимально может достигать значений порядка 120 мс. Полученные минимальные значения несколько ниже упомянутых ранее 30 мс за счет возможности отнесения части релаксационных колебаний к самому звуку, а не к смычке.

Аналогичной не менее важной для корректной работы VAD-алгоритма информацией является минимально возможная длительность звука. Минимальная длительность активной речи достигается для одиночных звуков, обрамленных с обеих сторон смычкой и паузой, и составляет порядка 30-35 мс: такую минимальную длительность имеют взрывные звуки, находящиеся в конце слова.

2.3.2 Средняя мощность звука, нормированная сумма модулей отсчетов, энергия

Энергия фрагмента РС является наиболее очевидным индикатором наличия/отсутствия в нем вокализации: как правило, энергия вокализованных фрагментов в несколько раз превышает энергию таких же по длительности невокализованных фрагментов РС.

Энергию имеет смысл сравнивать для равных по длительности фрагментов сигнала (т.е. при оконной обработке). Однако в случае изучения характеристик целых фонем необходимо использовать параметры, усредненные по времени. Для энергии соответствующим параметром является средняя мощность сигнала. Средняя мощность рассчитывается нормированием энергии фрагмента сигнала на его длительность (объем выборки). Так как исследуемые фонограммы получены в одном сеансе записи и по уровню нормированы к значениям *±1*, то и вычисляемая средняя мощность является относительной величиной:

$$P_{cp} = \frac{\sum_{i=1}^{N} x_i^2}{N}, \tag{2.3}$$

где x_i – i-й элемент выборки, N – объем выборки.

Для упрощения процесса вычисления для аналогичных целей можно использовать функцию суммы модулей отсчетов фрагмента сигнала (MSF, Magnitude Sum Function) [40], также нормированную к числу отсчетов на рассматриваемом интервале:

$$MSF = \frac{\sum_{i=1}^{N} |x_i|}{N}. \tag{2.4}$$

Общее распределение средних мощностей по фонемам показано на рисунке 2.7. В данном случае наблюдается еще больший – по сравнению с оценками средних длительностей фонем – разброс возможных значений средней мощности звука для одинаковых фонем.

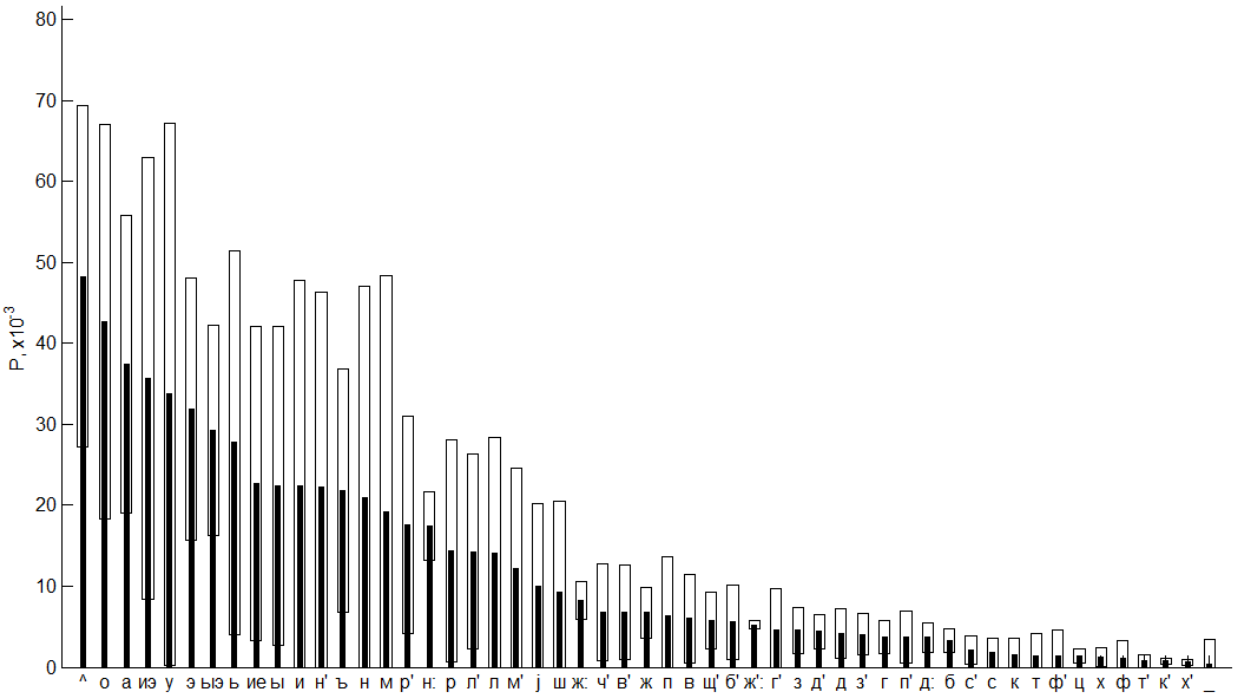


Рисунок 2.7 – Оценки средних мощностей звуков ±σ-размахи

Наибольшей средней мощностью обладают гласные звуки – это напрямую следует из механизма образования гласных, которому соответствует отсутствие в речевом аппарате препятствий для свободного прохождения воздуха. Наиболее мощными являются звуки [а], [о], которые встречаются в большинстве случаев только под ударением, и звук [^], соответствующий буквам А, О, Э в первой слабой позиции (т.е. в предударном слоге), а также во второй слабой позиции в абсолютном начале слова. Соответствующие варианты звуков для второй слабой позиции фонемы – [ъ] и [ь] – уже в среднем имеют в два раза меньшую мощность.

Таблицы с полученными средними мощностями для ударных и безударных гласных показаны в Приложении Б: таблицы Б.2 и Б.3 соответственно. Мощность звука в безударной позиции уменьшается обычно не более чем в полтора раза от мощности в сильной позиции.

Энергия реализаций фонем рассчитывается по формуле, аналогичной формуле (2.3), но без нормировки на длительность (количество отсчетов) звука. Результаты вычисления средних энергий звуков приведены в виде диаграммы в Приложении Б на рисунке Б.1.

Важно отметить, что наибольшую энергию несут ударные гласные звуки. В свою очередь, наименьшей энергией обладают короткие взрывные звуки и более длительные, но обладающие меньшей средней мощностью, шумные звуки. Твердый [т] несет в среднем в 200-300 раз меньше энергии, чем ударный [о].

2.3.3 Частота переходов через нуль

Параметр *ZCR* частоты пересечений сигналом нулевого уровня, наряду с энергией, является простым (но не достаточным) индикатором вокализации звука. Для вокализованных звуков в силу присутствия колебаний основного тона частота пересечения относительно небольшая. В то время как для невокализованных звуков присутствие шумовой составляющей в звуках приводит к большим значениям параметра.

Частота переходов через нуль может быть вычислена по формуле:

$$ZCR = \frac{\sum_{i=1}^{N} \left| sign(x_i) - sign(x_{i-1}) \right|}{N}, \tag{2.5}$$

где *sign(x)* – функция знака, принимающая значения *±1*, в зависимости от знака операнда *x*.

Параметр частоты пересечений нуля основное применение находит в алгоритмах сегментации более низкого, нежели сегментация по фонемам, уровня. В частности, этот параметр применяется для сегментации на фрагменты «вокализованный/невокализованный».

Результаты вычисления обобщенного по дикторам среднего значения частоты пересечений нуля приведены в таблице Б.4 Приложения Б и в виде диаграммы визуализированы на рисунке 2.8. Видно, что, действительно, для вокализованных звуков частота переходов через нуль значительно ниже, чем для шумных звуков. Однако также видно, что использование одного только этого параметра не может дать однозначного ответа о наличии или отсутствии в некотором фрагменте РС факта вокализации. В частности, вокализованные звуки

[з], [ж] и их палатализованные вариации содержат значительную шумовую составляющую, поэтому по параметру частоты пересечений нуля они могут попадать в класс чисто шумных звуков. Среди вокализованных звуков значение *ZCR* меньше у обладающих более простой структурой периода ОТ – см. ниже пункт 2.3.5 «Количество переколебаний на одном периоде основного тона».

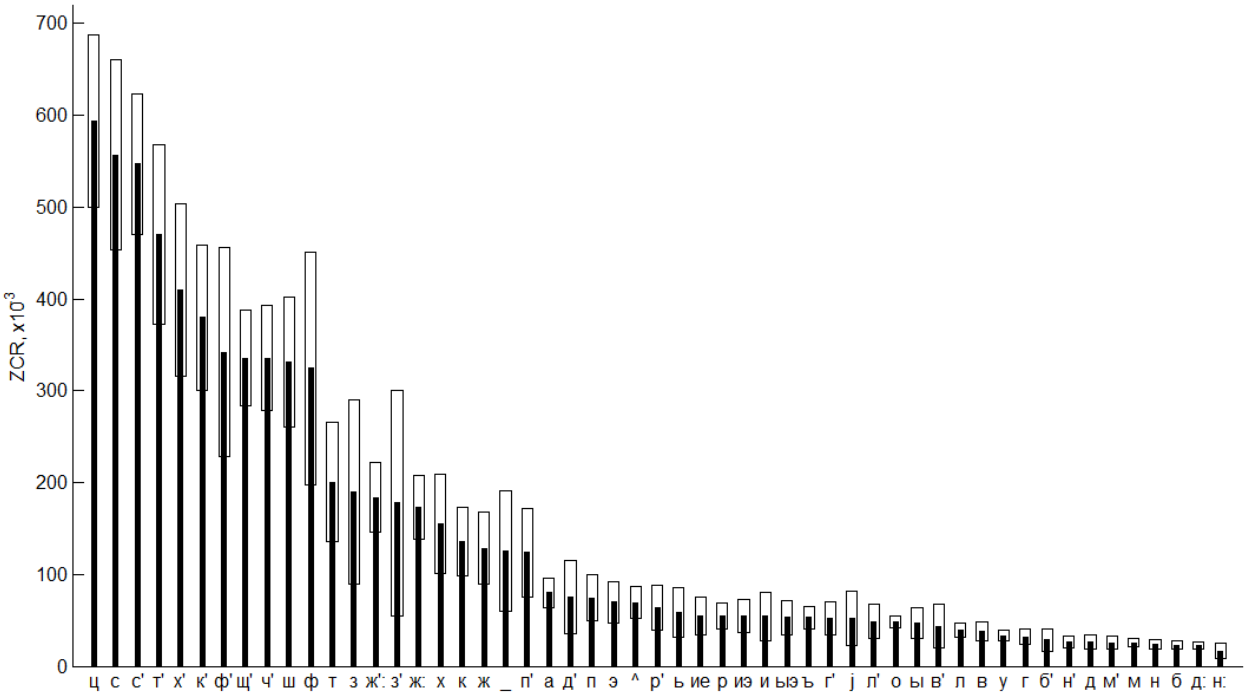


Рисунок 2.8 – Оценки средних частот пересечений нуля для различных звуков ±σ-размахи

2.3.4 Мел-частотные кепстральные коэффициенты (MFCC)

Использование мел-частотных кепстральных коэффициентов в задачах сегментации РС было рассмотрено выше в пункте 1.2.2 «Кепстральный анализ».

В данном исследовании вычисление векторов мел-частотных кепстральных коэффициентов осуществлено посредством соответствующей программной функции, включенной в расширение VoiceBox для MATLAB. При вычислениях сигнал разбивается на окна оценивания по 512 отсчетов с коэффициентом перекрытия 0,5 и взвешиванием окном Хэмминга. В мел-частотной области

используются 30 треугольных фильтров (примерно 2,1 на октаву), и, в итоге, вычисляется вектор из 12-ти MFCC-коэффициентов (не включая нулевой).

На рисунках 2.9–2.11 показаны последовательности MFCC-коэффициентов для звуков [А], [О] и [с] (гласные ударные). Звуки расположены внутри – то есть не в абсолютном начале и не в абсолютном конце – слов, взятых из перечня слов для исследования и произнесенных одним диктором-женщиной. На каждом графике показана выборка из десяти векторов MFCC-коэффициентов. На рисунке 2.12 на одном графике отображены коэффициенты, полученные при произнесении ударного [А] двумя дикторами – женщиной и мужчиной.

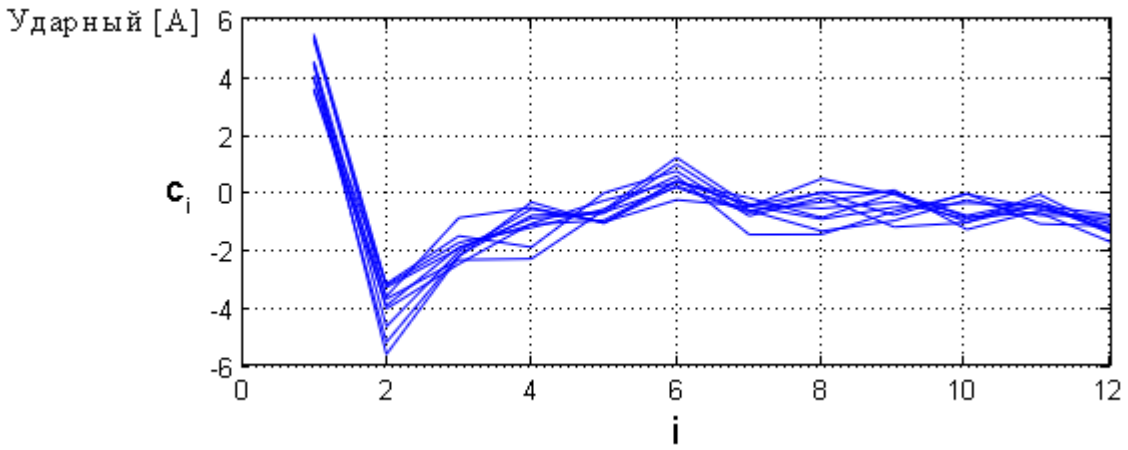


Рисунок 2.9 – Визуализация выборки векторов MFCC-коэффициентов для ударного звука [А]

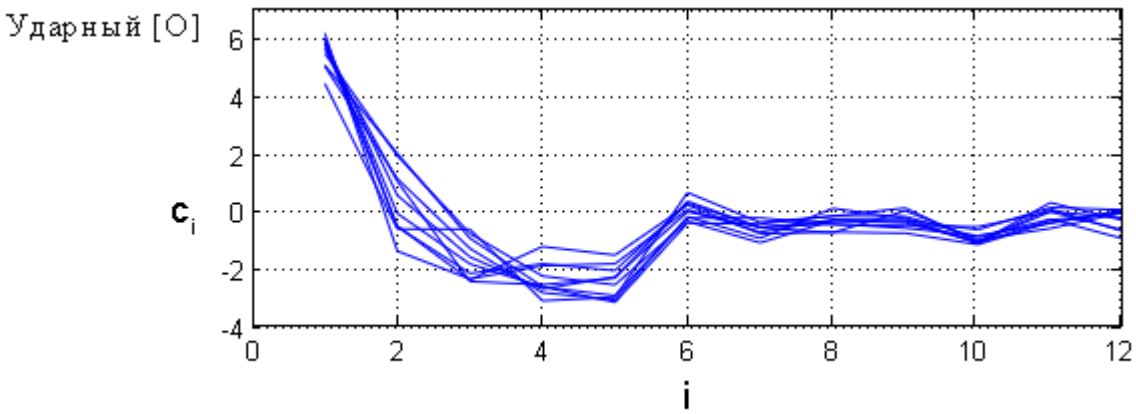


Рисунок 2.10 – Визуализация выборки векторов MFCC-коэффициентов для ударного звука [О]

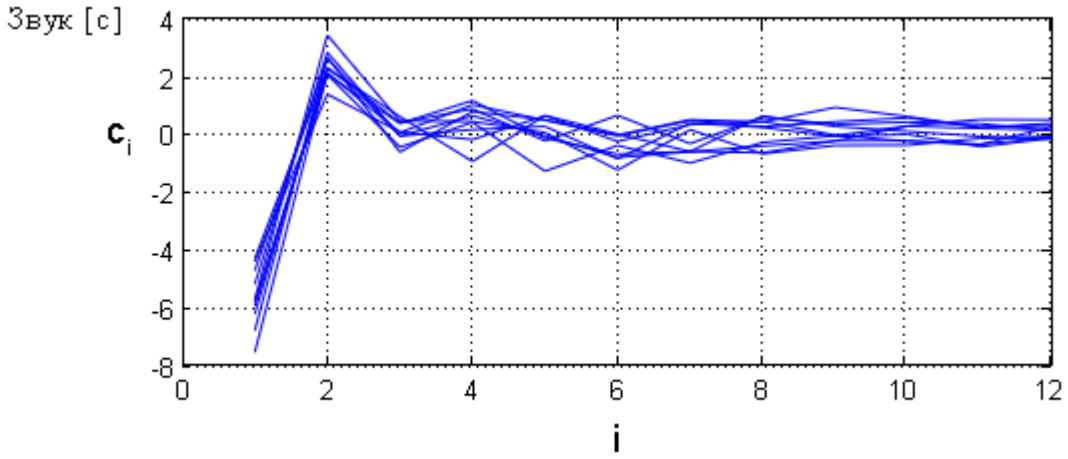


Рисунок 2.11 – Визуализация выборки векторов MFCC-коэффициентов для звука [c]

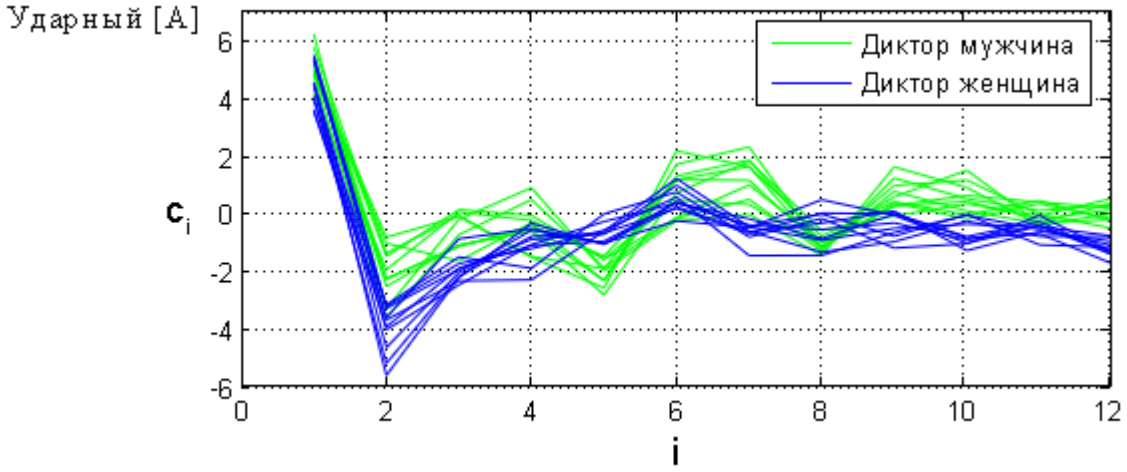


Рисунок 2.12 – Визуализация выборки векторов MFCC-коэффициентов для ударного звука [á], произнесенного двумя дикторами

По приведенным рисункам видно, что для одинаковых звуков линии, представляющие вектора MFCC-коэффициентов, выстраиваются в единую закономерность. В то же время для разных звуков такие закономерности различаются. Это делает мел-частотные кепстральные коэффициенты удобным инструментом при анализе РС, что объясняет их широкое применение в приложениях автоматической обработки речи.

В ходе исследования были получены вектора MFCC-коэффициентов для всех имеющихся в базе реализаций звуков и вычислены усредненные значения коэффициентов для всего перечня фонем. Это позволило успешно использовать MFCC в комплексном алгоритме сегментации, рассматриваемом ниже в

подразделе 3.7 «Многопараметрические алгоритмы многоуровневой временной сегментации речевых сигналов».

2.3.5 Количество переколебаний на одном периоде основного тона

Во временном представлении сигнала совокупность значений формантных частот находит отражение в количестве переколебаний или, иными словами, количестве низкочастотных минимумов (или, что эквивалентно, максимумов).

Как показано ниже в подразделе 2.5 «Исследование особенностей основных классов звуков русской речи», для вокализованных звуков с более простой структурой строения низкочастотной составляющей периодов ОТ характерным является значительное разнесение по частоте первой и высших формант.

Общая картина значений данного параметра, полученная для вокализованных звуков с обобщением по дикторам и позициям звука, представлена на рисунке 2.13.

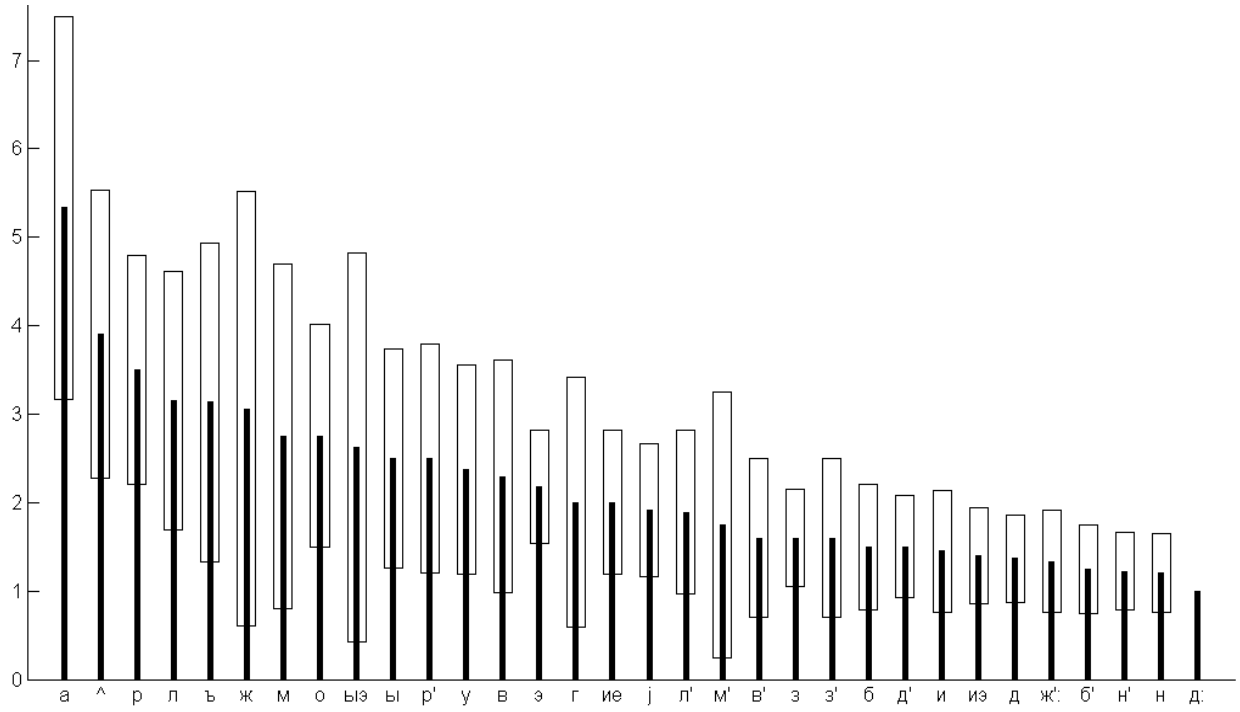


Рисунок 2.13 – Диаграмма распределения среднего количества переколебаний на периоде ОТ по звукам ±σ-размахи

Важно заметить, что количества переколебаний на периоде ОТ зависят не только от произносимого звука, но и от характеристик речевого аппарата диктора,

поэтому целесообразно построить диаграммы, аналогичные показанной на рисунке 2.13, отдельно для каждого диктора. Такие диаграммы получены и приведены в Приложении Б на рисунках Б.2 и Б.3. При раздельном рассмотрении дикторов наблюдается, во-первых, значительно меньший разброс значений параметра по фиксированным фонемам, а во-вторых, закономерность для всех звуков, заключающаяся в большем количестве переколебаний у диктора-мужчины. Данный факт объясняется различием частот ОТ дикторов разных полов: у женского голоса частота колебаний связок достаточно высокая, поэтому первая гармоника ОТ превалирует в формировании первой форманты; в то же время для мужского голоса первая форманта в большей степени затрагивает вторую-третью гармоники частоты ОТ (см. ниже в подразделе 2.5 «Исследование особенностей основных классов звуков русской речи»).

На основе полученных диаграмм можно выделить ряд звуков, которые, независимо от диктора, будут иметь «сложную» структуру периода основного тона, то есть более одного переколебания. В первую очередь, к таким звукам можно отнести звук [а] и [^]. Здесь также можно отметить убывание среднего числа переколебаний с убыванием «силы позиции» гласного: максимальное значение для [а] (сильная позиция), меньше для [^] (первая слабая позиция) и еще меньше для [ь] и [ъ] (вторая слабая позиция). Аналогичным образом среди согласных звуков присутствует ряд звуков с простой структурой, по форме близкой к гармоническому колебанию: [б], [н] и ряд других звуков. Стоит также отметить, что представленная картина может незначительно измениться при изменении алгоритма вычисления переколебаний. В частности, можно получить больший процент единичных значений для квазигармонических звуков, если не учитывать незначительные внутренние переколебания. В качестве примера звука, в котором алгоритмом было выявлено два переколебания на периоде ОТ при простой структуре низкочастотной составляющей, можно показать одну из реализаций звука [д] (фонограмма слова «*Дал*» диктора-мужчины), рисунок 2.14.

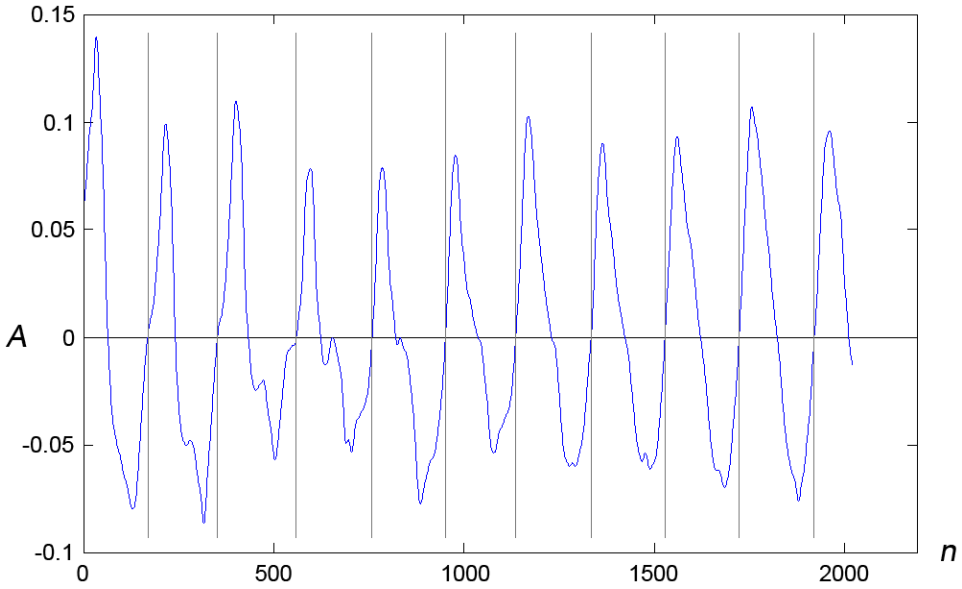


Рисунок 2.14 – Реализация звука [д] с неоднозначным определением числа переколебаний на одном периоде ОТ

Таким образом, параметр количества переколебаний НЧ-составляющей сигнала на периоде ОТ определяется, в первую очередь, частотой колебания голосовых связок и соотношением частот первой и высших формантных частот в рассматриваемом фрагменте сигнала. Параметр является в значительной степени дикторозависимым.

2.4 Разработка таксономии звуков русской речи с точки зрения задачи сегментации

Проведенное исследование позволяет сформировать иерархическую структуру групп (таксономию) исследуемых аллофонов русской речи с точки зрения существенных для задачи сегментации сигнальных признаков.

Наиболее крупное деление осуществляется по признаку наличия в звуке вокализации: можно выделить звуки вокализованные (звонкие) и невокализованные (глухие).

В вокализованные звуки входят гласные и согласные, это разделение генетически обусловлено наличием в речевом тракте преграды при произнесении согласных [107]. Как следствие, гласным характерна большая по сравнению с согласными средняя мощность (однако этот параметр недостаточен для однозначного отнесения звука к гласным или согласным, см. рисунок 2.7).

Среди вокализованных согласных в русском языке существует несколько звуков со значительной шумовой составляющей, снижающей точность VAD-сегментации – звуки, соответствующие аллофонам [ж], [ж:], [ж’:], [з] и [з’]. Это щелевые звонкие звуки, при произнесении которых в речевом аппарате возникает высокая турбулентность. Остальные вокализованные согласные произносятся при меньших препятствиях в речевом аппарате, вследствие чего в их временной структуре вокализованная составляющая преобладает над шумовой.

Наконец, гласные звуки и малошумные вокализованные согласные могут быть подразделены на звуки с простой и сложной структурой ОТ. Данное деление основывается на количестве переколебаний в структуре НЧ-компоненты периода ОТ и проявляется в чистом виде для голосов с высоким тоном (для женских). У низких голосов для всех вокализованных звуков количество переколебаний становится больше (см. рисунки Б.2 и Б.3 в Приложении Б). Сложная структура периода основного тона является значимой причиной снижения надежности алгоритмов ОТ-сегментации.

Отдельного внимания заслуживает единственный по фонологической классификации дрожащий звук [р]. Данный звук является многоударным [108], эта особенность хорошо прослеживается на сонограммах как вертикальные разрывы на спектральной картине (см. рисунок 2 в [108]). К слову, дрожащие звуки отсутствуют в английском языке – одном из самых распространенных языков, на который ориентирована значительная часть современных алгоритмов обработки РС.

В группе невокализованных звуков присутствуют только согласные. При этом среди невокализованных достаточно явно выделяются две подгруппы: взрывные и шумные.

Отнесенные в настоящей классификации к взрывным звуки [к], [к’], [п], [п’], [т] имеют очень короткую длительность по сравнению со всеми остальными звуками – на рисунке 2.6 это единственные 5 звуков, имеющие средние длительности меньше средней длительности паузы-смычки. С точки зрения фонетики к глухим взрывным звукам относятся [к], [п], [т] и соответствующие им

палатализованные звуки. Однако в предлагаемой классификации [т’] отнесен к подгруппе шумных в силу закономерного отсутствия характерного для взрывных звуков резкого амплитудного скачка и сравнительно большой длительности – ср. форму и длительность звуков [т] и [т’] на рисунке 2.15. Также отдельно стоит выделить звук [п], имеющий значительную низкочастотную составляющую и наименьшую из всех звуков среднюю длительность.

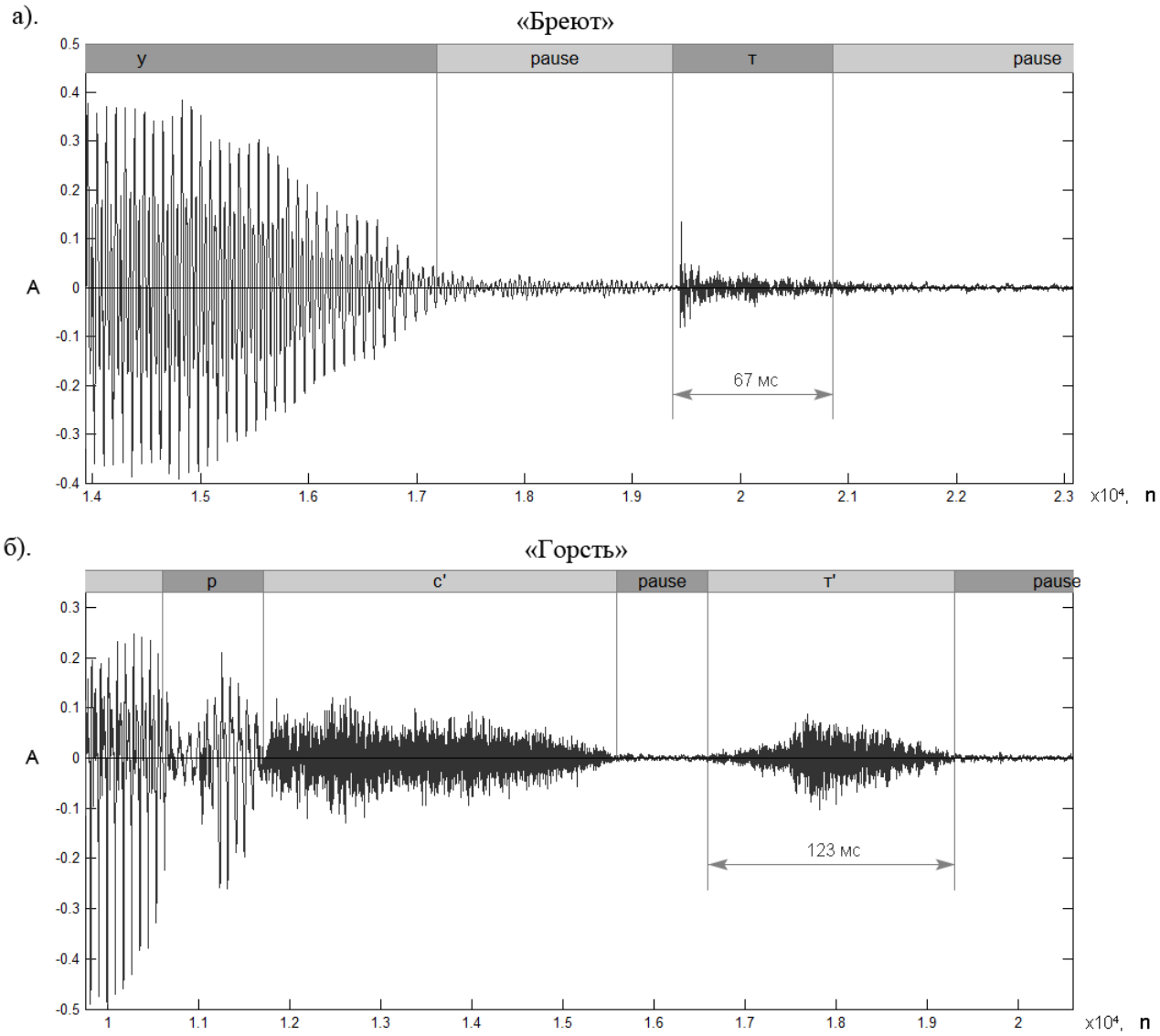


Рисунок 2.15 – Произнесенные одним диктором звуки: а) твердый [т] в конце слова, б) мягкий [т'] в конце слова

К невокализованным шумным, помимо [т’], отнесены аффрикаты и глухие спиранты. Эти звуки образуются при значительной близости артикуляторов, им

характерна выраженная шумовая структура и большая, нежели у невокализованных взрывных, длительность.

Получаемая в итоге иерархическая структура групп звуков русской речи показана на рисунке 2.16.

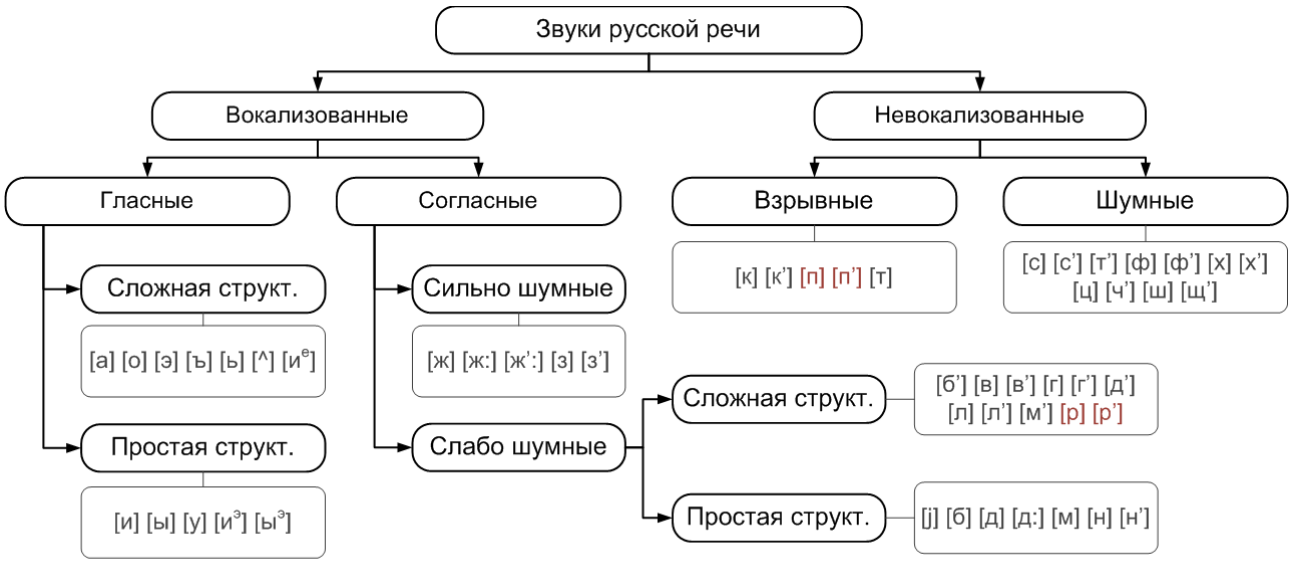


Рисунок 2.16 – Таксономия звуков русской речи по существенным для задачи сегментации признакам

Важно еще раз подчеркнуть, что данная классификация основана на различиях групп звуков, представленных в виде акустических сигналов, важных с точки зрения задачи сегментации – это обуславливает наличие расхождений с классификациями, формируемыми с точки зрения фонетики русского языка (см. рисунки 1.3, 1.4, 1.5).

2.5 Исследование особенностей основных классов звуков русской речи

Ряд особенностей реализаций фонем, их взаимного влияния достаточно сложно выявить автоматически на основе имеющейся фонемной сегментации базы исследуемых фонограмм. Это связано, во-первых, с отсутствием однозначных достоверных алгоритмов вычисления ряда параметров РС (к примеру, алгоритма вычисления значений формантных частот), а во-вторых, с

глубоким взаимодействием соседних звуков, которое затруднительно учитывать автоматически при раздельном изучении фонем.

В данном подразделе проведен анализ спектральных картин классов звуков согласно предложенной выше классификации.

2.5.1 Вокализованные гласные

Для вокализованных гласных в силу активности голосовых связок (генераторная часть голосового аппарата) и отсутствия препятствий для прохождения воздуха спектр принимает ярко выраженную формантную структуру, а также характерен присутствием гармоник частоты основного тона.

При этом гласным, отнесенным в таксономии к вокализованным простой структуры, характерна выраженная первая форманта и менее выраженная и сравнительно далеко отнесенная по частоте вторая форманта.

Для гласных же со сложной НЧ-структурой периода ОТ вторая форманта выражена хорошо и по частоте расположена ближе к первой. Характерные спектрограммы вокализованных гласных показаны на рисунке 2.17.

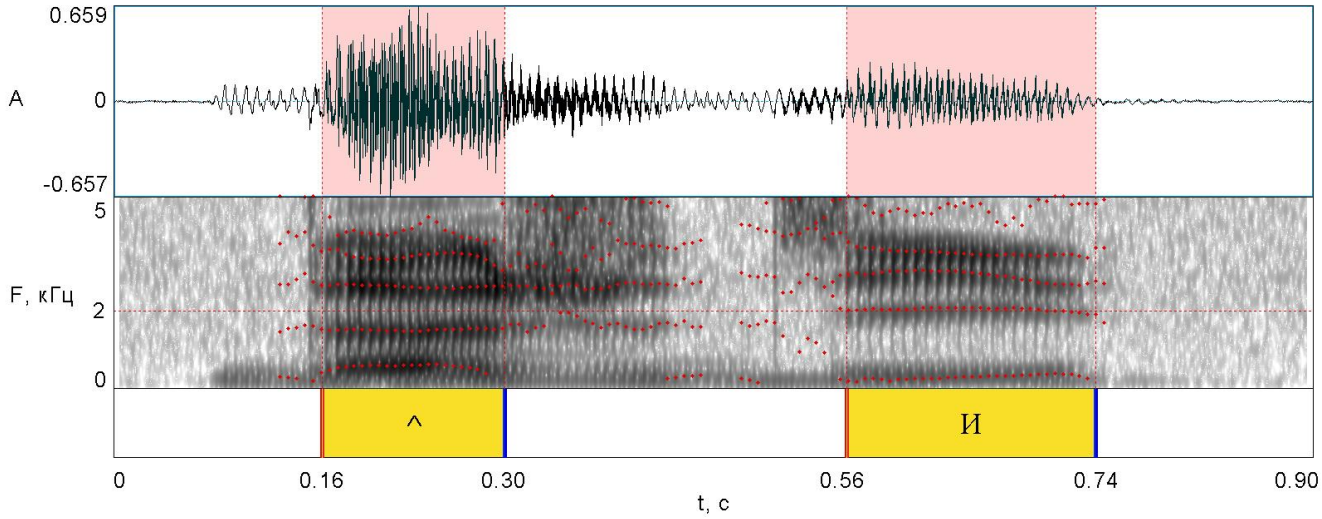


Рисунок 2.17 – Две вокализованные гласные слова «Дожди», мужской голос: гласный сложной структуры [^] и гласный простой структуры [И]

2.5.2 Вокализованные согласные

Для вокализованных согласных спектр также во многом имеет формантную структуру с присутствием гармоник частоты основного тона, однако в данном случае форманты выше второй значительно менее выражены и, порой, трудноразличимы. При этом для подгруппы сильно шумных звуков на форманты дополнительно накладывается широкополосная шумовая составляющая, выраженная на частотах от 2,5 кГц и выше. Для подгрупп согласных простой и сложной НЧ-структур периода ОТ в целом сохраняются закономерности, характерные для соответствующих подгрупп рассмотренных выше вокализованных гласных.

Ряд звуков содержит свои уникальные отличительные особенности. К примеру, для мягкого [л’], а также для [р] в начале звука характерно общее краткое падение всех спектральных составляющих, кроме первых гармоник (соответствуют первой форманте), и, изредка, второй форманты.

Мягкий [д’] значительно изменяется в процессе произнесения, начинаясь с чистых колебаний первой форманты (первые 1-3 гармоники основного тона) и продолжаясь появлением шумовой составляющей от 4 кГц во второй части звука.

Примеры характерных спектрограмм данной группы звуков показан на рисунках 2.18 и 2.19.

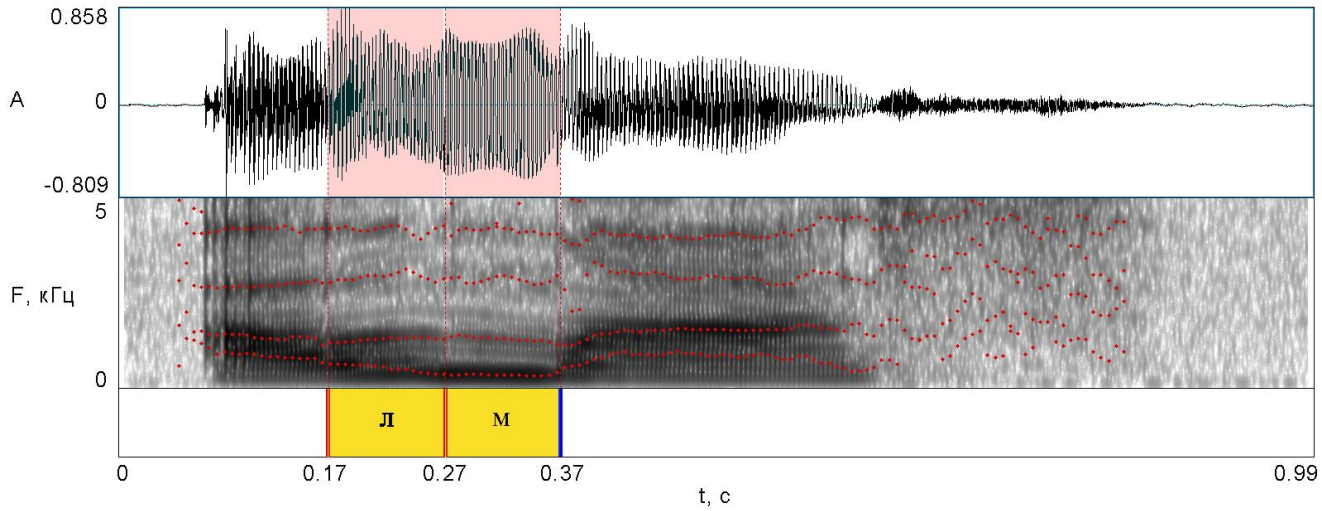


Рисунок 2.18 – Две вокализованные слабошумные согласные слова "Алмаз": сложной структуры [л] и простой структуры [м], женский голос

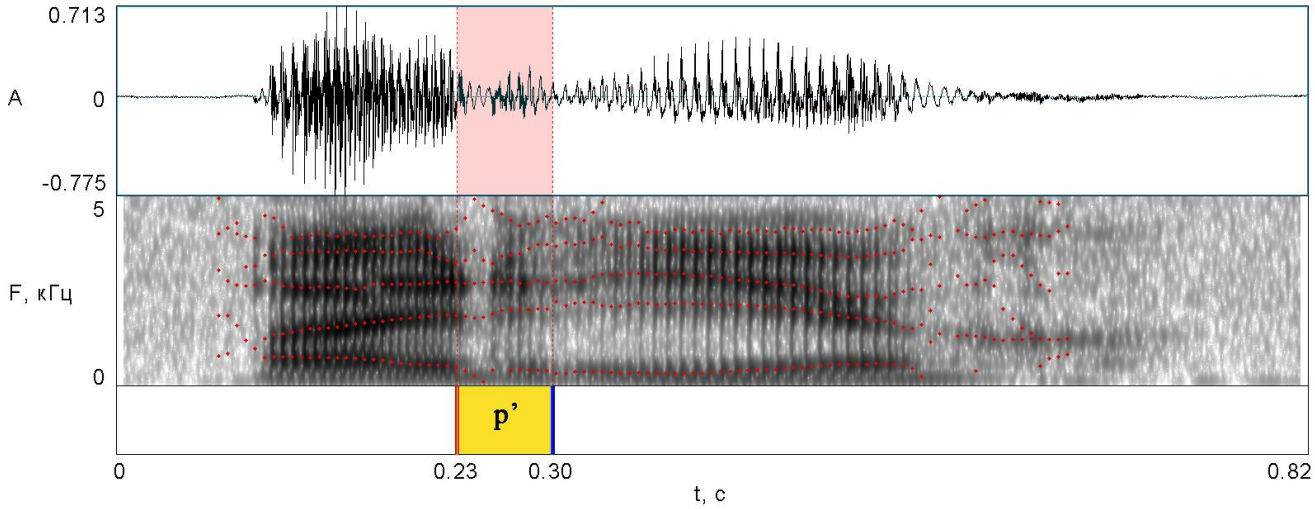


Рисунок 2.19 – Вокализованный слабошумный согласный звук [р'] слова "Орех", мужской голос

2.5.3 Невокализованные взрывные

Данная группа звуков характерна малой длительностью, а также наличием одного-двух кратковременных взрывов, вызываемых резким раскрытием препятствия в речевом тракте и выражающихся в виде кратковременного повышения интенсивности сигнала по всему диапазону частот.

В связи с высокой скоростью изменения характерных особенностей сигнала для спектрального рассмотрения данных звуков следует выбирать значительно более короткие окна фурье-анализа либо применять иные методы исследования.

Здесь следует отметить роль резонансных полостей речевого аппарата, определяющих формантные частоты: в спектрах взрывных звуков усиления интенсивности в формантных областях заметны, причем сильнее для тех звуков, место образования которых находится глубже, например, для заднеязычного [к]; и, наоборот, слабее для звуков, образующихся на выходе речевого аппарата, например, для [п], у которого местом образования взрыва являются губы.

Перечисленные особенности хорошо заметны в спектральном представлении слова «Экспорт», отображенном на рисунке 2.20.

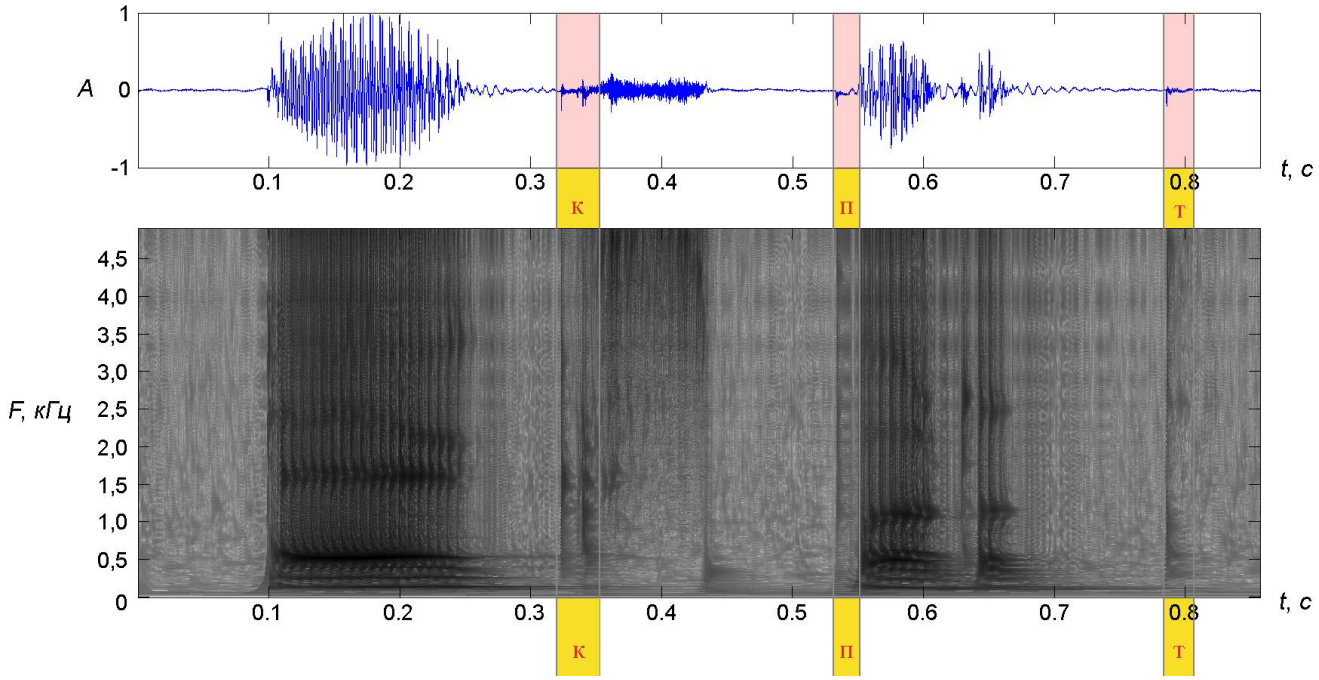


Рисунок 2.20 – Ряд реализаций взрывных звуков на примере слова «Экспорт», мужской голос

2.5.4 Невокализованные шумные

Невокализованные шумные характерны отсутствием гармоник частоты основного тона (за исключением областей взаимного проникновения с вокализованными звуками), а также наличием шумовой составляющей, начинающейся с частот от 2 кГц. У звука [с’], в сравнении с другими звуками данной группы, шумы более интенсивны в области 5 кГц.

Мягкий [т’], относящийся с фонетической точки зрения к взрывным, в спектральной картине, действительно, может содержать (но не всегда достаточно выраженный) характерный для своего твердого аналога кратковременный широкополосный взрыв.

При произнесении данных звуков голосовые связки неактивны, поэтому формантные области спектра выражены слабо. При этом, как и при произнесении взрывных согласных, степень выраженности формант определяется местом образования звука. Характерным при этом является заднеязычный [х], в спектрограммах реализаций которого особенно выражена вторая форманта (рисунок 2.21).

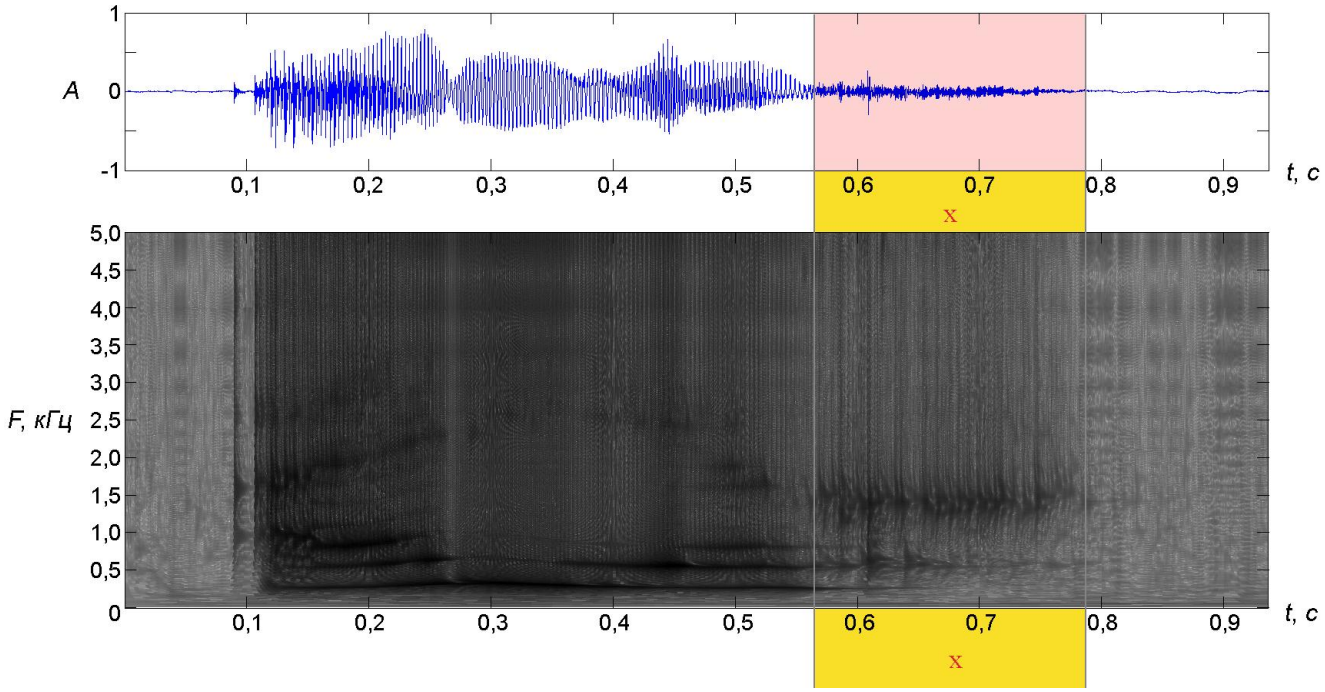


Рисунок 2.21 – Невокализованный шумный заднеязычный звук [х] на примере реализации слова «Орех», женский голос

2.6 Основные выводы по разделу

Наиболее технически сложным уровнем временной сегментации является фонемная сегментация. В русском языке 43 общепринятые фонемы, однако, при речеобразовании, в зависимости от взаимного расположения фонем, их положения относительно границ слова, ударного звука, фонемы реализуются в виде различных своих аллофонов. Полный список аллофонов крайне широк, кроме того большинство аллофонов одной фонемы настолько похожи, что могут не быть различены на слух даже носителями языка. Вследствие этого более целесообразным при фонемной сегментации является применение укрупненной классификации, содержащей только существенные аллофоны фонем.

В исследовании рассматриваются 52 существенных аллофона фонем русского языка, кроме того, отдельной фонемой рассматривается пауза-смычка. Результаты исследования хранятся в специально разработанной базе данных, поэтому становится возможным проведение любого углубления классификации комбинаторных и позиционных аллофонов.

В зависимости от конфигурации речевого аппарата при произнесении тех или иных звуков в речевом сигнале наблюдаются характерные особенности, обуславливающие возможные уровни временной сегментации. В частности, могут выделяться сегменты пауз, активности речи, вокализованных, взрывных, шумных звуков, периодов колебаний основного тона и т.д.

Знания об основных правилах чтения слов и словарь ударений русского языка позволяют реализовать методы автоматизации транскрибирования слов в последовательность аллофонов. Этот подход позволяет не только упростить проведение исследования и сделать его результаты более надежными, но и является основой для построения контекстно-зависимых алгоритмов временной сегментации РС.

Изучение значений сигнальных параметров реальных реализаций звуков позволяет получить статистические данные, необходимые для успешного создания алгоритмов сегментации различного уровня: начиная от сегментации «речь/пауза» (важен диапазон изменения длительности смычки, статистика по длительности непрерывной речевой активности) и заканчивая фонемной сегментацией (статистические данные о параметрах фонемы в зависимости от ее положения в слове, относительно других фонем, относительно ударного гласного). Выделение общих сигнальных признаков для разных групп аллофонов позволяет сформировать иерархическую структуру групп звуков русской речи, отражающую основные уровни сегментации и нюансы эффективности VAD и ОТ-сегментации.

3 РАЗРАБОТКА АЛГОРИТМОВ СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ И СМЕЖНЫХ АЛГОРИТМОВ

3.1 Системный подход к сегментации

3.1.1 3 базовых уровня сегментации

На основе информации о строении РС, об особенностях реализаций звуков русской речи задача временной сегментации фонограмм может быть разбита на несколько последовательно осуществляемых этапов, каждый из которых представляет собой соответствующий уровень сегментации (рисунок 3.1).

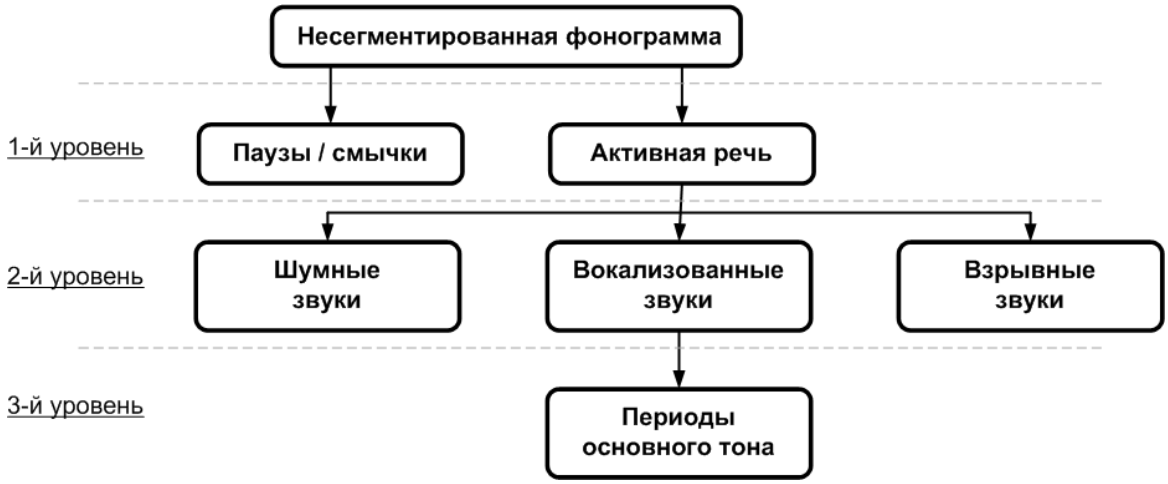


Рисунок 3.1 – Три уровня временной сегментации РС

На первом уровне сегментации происходит разбиение речевого сигнала на фрагменты активной речи и пауз (применяется VAD алгоритм). При этом в роли паузы может выступать и смычка (см. рисунок 2.3).

Вторым уровнем сегментации является разбиение активных участков речи на фрагменты, соответствующие трем основным типам звуков: вокализованные, шумные и взрывные.

Третий уровень сегментации относится только к вокализованным фрагментам и заключается в их подсегментации на отдельные периоды ОТ. Следует отметить, что разбиение на периоды ОТ необязательно является финальной стадией сегментации: в дальнейшем на ее основе может быть выполнено разбиение вокализованных фрагментов на аллофоны / группы из

однотипных аллофонов. В данном случае становится возможным построение собственных нейронных сетей, марковских моделей и других эталонов для различных типов фонем. Так, в системе верификации, разрабатываемой компанией SPIRIT Corp. [109] процедура классификации, основанная на смешанных гауссовых моделях, позволяет принимать во внимание отдельные фонемы или группы фонем и особенности их произнесения конкретным диктором.

3.1.2 Структура обобщенного алгоритма сегментации

На рисунке 3.2 показан предлагаемый системный подход к сегментации и последующей параметризации РС. Первые шаги обработки (блоки 1...3) соответствуют работе VAD-алгоритма. Основным информативным параметром (блок 9) сегментов паузы является их длительность.

Рисунок 3.2 – Функциональная схема системного подхода к сегментации и параметризации РС

Над участками активной речи производится следующий, второй, уровень сегментации, на котором выделяются сегменты, соответствующие типам звуков: вокализованные (блок 4), шумные (блок 8), взрывные глухие (блок 7).

При параметризации речевых сегментов разных типов могут использоваться как различные наборы параметров, так и единые преобразования фрагментов РС для формирования однотипных параметров (например, использование для параметризации спектрального преобразования РС, формирование корреляционной функции и ее параметризация, и т.д.).

Возможные комбинации разных граничащих типов звуков, включая паузу:

- пауза-шумный (например, «шел»),
- шумный-пауза (например, «пустошь»),
- пауза-взрывной (например, «палка»),
- взрывной-пауза (например, «окоп»),
- пауза-вокализованный (например, «аорта»),
- вокализованный-пауза (например, «аорта»),
- шумный-взрывной (перед взрывным звуком присутствует смычка, например, «штопор»),
- взрывной-шумный (например, «псевдоним»),
- шумный-вокализованный (например, «шар»),
- взрывной-вокализованный (например, «пар»),
- вокализованный-взрывной (перед взрывным звуком присутствует смычка, например, «суп»; смычка может содержать слабые релаксационные колебания предваряющего вокализованного звука).

Вокализованные фрагменты подвергаются дополнительной сегментации на периоды ОТ с последующей параметризацией и анализом трендов и разладок (блоки 5, 12, 14...16).

В зависимости от применения результатов автоматической сегментации в конкретных функциональных алгоритмах, может выполняться дополнительное разбиение однотипных звуковых фрагментов РС на отдельные фонемы.

В примере выполнения автоматической сегментации на фрагменты неравной длительности для задачи распознавания (см. раздел 4 «Приложения разработанных алгоритмов многоуровневой временной сегментации РС»), общая последовательность работы алгоритмов сегментации должна быть следующей:

а. детектирование типов сегментов РС (классификация фрагментов РС на паузы, шумные, взрывные, вокализованные);

б. уточнение временных границ сегментов (алгоритм повторного прохода)

в. параметризация сегментов РС

г. анализ и корректировка состава сегментов РС (детальный анализ сегментов с возможным разделением на подсегменты, соответствующие раздельным фонемам);

д. классификация выделенных фонем.

3.1.3 Метод сравнения эффективности работы однотипных алгоритмов сегментации

Один из основных вопросов при разработке алгоритмов сегментации: является ли новый алгоритм более эффективным по сравнению с уже существующими? От ответа на этот вопрос зависит целесообразность применения разработанного алгоритма в речевых системах. Далее предлагается систематизированный подход для количественного сравнения точности однотипных алгоритмов сегментации РС.

По результатам параметризации для каждого небольшого (вплоть до одного отсчета – сэмпла) фрагмента речевого сигнала вычисляется количественная оценка некоторого параметра (к основным методам параметризации можно отнести: линейное предсказание, кепстральный анализ, вейвлет-преобразование, анализ спектра модуляции [110]). Таким образом, вектор параметров представляет собой набор численных оценок с обязательным соотнесением каждой отдельной оценки к соответствующему временному фрагменту сигнала.

Задача сегментации – выделение временных границ между имеющими определенные общие свойства фрагментами сигнала. В результате сегментации

могут быть получены либо временные метки границ сегментов без их характеризации, либо временные метки с указанием ограничиваемого ими типа сегмента, то есть с качественной характеризацией (например, VAD алгоритм).

Процессы параметризации и сегментации тесно связаны друг с другом. Сегментация сигнала осуществляется по предварительно оцененным параметрам. При этом для каждого применяемого параметра подбирается порог, и в каждый момент пересечения порога функцией зависимости параметра от времени производится сегментация, то есть членение речевого сигнала. Например, в реализации VAD-алгоритма может рассматриваться параметр средней на небольшом интервале времени мощности сигнала [111]. В свою очередь, в дальнейшем по результатам сегментации для сегментов разных типов вычисляются разные группы параметров. Например, для вокализованного сегмента может быть рассмотрен параметр длительности периода ОТ диктора, в то время как для шумных звуков и пауз такой параметр смысла не имеет. Для большего обобщения процессов сегментации и параметризации, сегментацию с качественной характеризацией типа сегмента можно рассматривать под термином «сегментация с качественной параметризацией» (в противовес классической численной параметризации) – в этом случае значением параметра является не число, а некоторый символьный код из алфавита типов сегментов.

Для оценки точности и надежности сегментации нет единого повсеместно используемого подхода, к тому же, создание такого подхода осложняется большим разнообразием возможных задач сегментации РС.

Для оценки эффективности VAD-алгоритмов в [112] и ряде других работ предлагаются следующие параметры:

- *FEC* (Front End Clipping) – ошибка определения границы перехода от паузы к речи;
- *MSC* (Mid Speech Clipping) – ошибочное определение речи как паузы;
- *OVER* – ошибка определения границы перехода от речи к паузе;
- *NDS* (Noise Detected as Speech) – ошибочное определение паузы как речи.

Здесь следует обратить внимание, что в данном методе анализа эффективности делается различие между ошибками определения границ сегментов (*FEC* и *OVER*) и ошибками определения типа сегмента (*MSC* и *NDS*).

В работе [113] для оценки точности VAD-алгоритмов предлагаются несколько другие параметры: вводятся коэффициенты успеха *HR0* и *HR1*:

$$HR0 = \frac{N_{0,0}}{N_0^{ref}}, \quad HR1 = \frac{N_{1,1}}{N_1^{ref}}, \tag{3.1}$$

где $N_{0,0}$ и $N_{1,1}$ – количества верно определенных соответственно интервалов пауз и речи (при разбиении сигнала на интервалы окном малой длительности), а N_0^{ref} и N_1^{ref} – реальное количество интервалов пауз и речи по эталонной сегментации.

Как уже было упомянуто выше, задача оценки эффективности сегментации стоит не только для VAD-алгоритмов, но и для алгоритмов, осуществляющих другие виды сегментации. При этом возникающие ошибки сегментации могут иметь разную цену в зависимости от типов сегментов, которые ошибка затрагивает, их взаимного расположения. В литературе отсутствует единый подход к формулировке параметров точности сегментации, учитывающих указанные нюансы.

В рамках диссертационной работы предложен метод сравнения эффективности анализируемого алгоритма сегментации и эталонного, позволяющий в полной мере учесть многообразие возможных типов возникающих ошибок сегментации и, в то же время, дающий возможность упростить перечень параметров эффективности при наличии равнозначных типов ошибок. Основой оценки результата сегментации являются длительности временных интервалов с ошибочной классификацией фрагмента речи. При этом ошибка маркируется в зависимости от того, к какому типу сегмент относится согласно эталону и к какому типу он был отнесен анализируемым алгоритмом. Кроме того, отдельно рассматриваются случаи неверного определения границы сегмента (в коде ошибки присутствует символ подчеркивания «_» соответственно

слева или справа от обозначения направления ошибки) и неверного определения типа сегмента (символ «_» размещается в центре между символами обозначения направления ошибки).

Пример показан на рисунке 3.3: каждый тип сегмента имеет свое символьное обозначение. Фрагменты с ошибочной сегментацией помечены кодами соответствующих ошибок. Например, ошибка *_AB* обозначает, что согласно эталону фрагмент относится к сегменту типа *A*, но из-за ошибки определения левой границы сегмента был отнесен к сегменту типа *B*. Ошибка *B_C* будет обозначать, что вместо сегмента типа *B* фрагмент ошибочно отнесен к сегменту типа *C*.

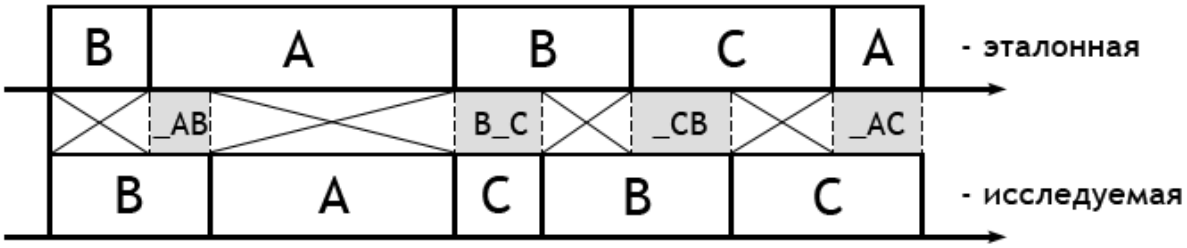


Рисунок 3.3 – Определение зон и маркировка ошибок при сравнении двух реализаций алгоритма: эталонной и исследуемой

Таким образом, можно получить совокупность коэффициентов ошибок, где каждый коэффициент, на примере маркировки *AB_*, определяется как:

$$AB\_ = \frac{T_{AB\_}}{T_A}, \tag{3.2}$$

где $T_{AB\_}$ – суммарная длительность всех ошибок типа *AB_*, T_A – суммарная длительность всех сегментов типа *A* согласно эталону.

При равной важности ошибок определения левой и правой границ сегмента они могут быть объединены в одну общую ошибку определения границ сегмента: *_AB_ = _AB + AB_*.

Если для системы неважно, ошибка произошла из-за неверного определения границ сегментов или из-за неверного определения типа сегмента, то

соответствующие коэффициенты ошибок могут быть объединены суммированием: *AB*=*_AB_* + *A_B*.

Кроме того, в ряде систем не делается различие по важности между направлением ошибки, то есть в таких системах можно объединять коэффициенты ошибок с зеркальными маркировками, например, *AB* и *BA*. При этом нормировка должна уже производиться по суммарной длительности обоих типов сегментов – и сегментов типа *A*, и сегментов типа *B*:

$$AB \,\&\, BA = \frac{T_{AB} + T_{BA}}{T_A + T_B}. \tag{3.3}$$

Как видно по формуле (3.3), при объединении ошибок с зеркальными маркировками необходима информация об относительной длительности каждого типа сегментов в эталонной фонограмме. Таким образом, объединенная ошибка *AB&BA* вычисляется через ошибки *AB* и *BA*:

$$AB \,\&\, BA = \frac{AB}{1+\frac{B_{len}}{A_{len}}} + \frac{BA}{1+\frac{A_{len}}{B_{len}}} = \frac{AB}{1+\frac{T_B}{T_A}} + \frac{BA}{1+\frac{T_A}{T_B}}, \tag{3.4}$$

где A_{len} и B_{len} для эталона длительностью L: $A_{len} = T_A / L;\ \ B_{len} = T_B / L$.

В простейшем случае для анализа эффективности алгоритмов сегментации все ошибки считаются равнозначными, и в качестве оценки эффективности алгоритма вычисляется одно единственное число – обобщенная ошибка. Если же все типы ошибок имеют различную важность, то перед объединением каждую элементарную ошибку следует домножить на коэффициент ее значимости. Для сохранения теоретического диапазона возможных значений суммарного коэффициента ошибки от *0* до *1* коэффициенты значимости ошибок следует также выбирать в пределах от *0* до *1*.

Ниже в подразделе 3.4.3 «Сравнение эффективности разработанных VAD-алгоритмов» в таблицах 3.7 и 3.8 приведены результаты сравнения по данной

методике эффективности двух разработанных вариантов реализации VAD-алгоритма между собой и с VAD-алгоритмом VoiceBox.

Необходимо также отметить, что в зависимости от конкретики задач, стоящих перед алгоритмами, может возникать потребность в иных подходах к оценке эффективности работы алгоритма сегментации. В частности, в работе [13] авторы рассматривают долю границ сегментов, определенных автоматически с точностью не менее *15* мс относительно эталонной сегментации. В исследовании [28] для оценки эффективности алгоритма вводятся три аналогичных параметра: P_0 – вероятность определения временного значения границы с погрешностью 0,01 с; P_1 – с погрешностью 0,02 с; P_2 – с погрешностью 0,03 с; $P_{>=3}$ – с погрешностью более 0,03 с; P_- – пропуск границ; P_+ – ложные границы сегментов.

3.2 Использование огибающей сигнала в алгоритмах сегментации

3.2.1 Алгоритм выделения огибающей речевого сигнала

Одной из важнейших составляющих временного анализа как фонограммы в целом, так и отдельных звуков, является рассмотрение характеристик огибающей (амплитудной модулирующей функции, МФ) сигнала/звука. В частности, реализация звука может иметь нерегулярную амплитудную модуляцию, и содержащаяся в огибающей информация будет утеряна при корреляционной или спектральной обработке. Одним из показательных примеров важности рассмотрения модулирующей функции может служить звук [р], который всегда имеет резко вогнутую составляющую огибающей (рисунок 3.4).

Во многих задачах обработки речи, использующих результаты сегментации, важна точность сегментации вплоть до одного колебания основного тона. Поэтому алгоритмы выделения огибающей, основанные на фильтрации сигнала или его преобразований, не всегда будут применимы в задачах сегментации из-за усредненности результата, то есть отсутствия привязки к конкретным отсчетам сигнала.

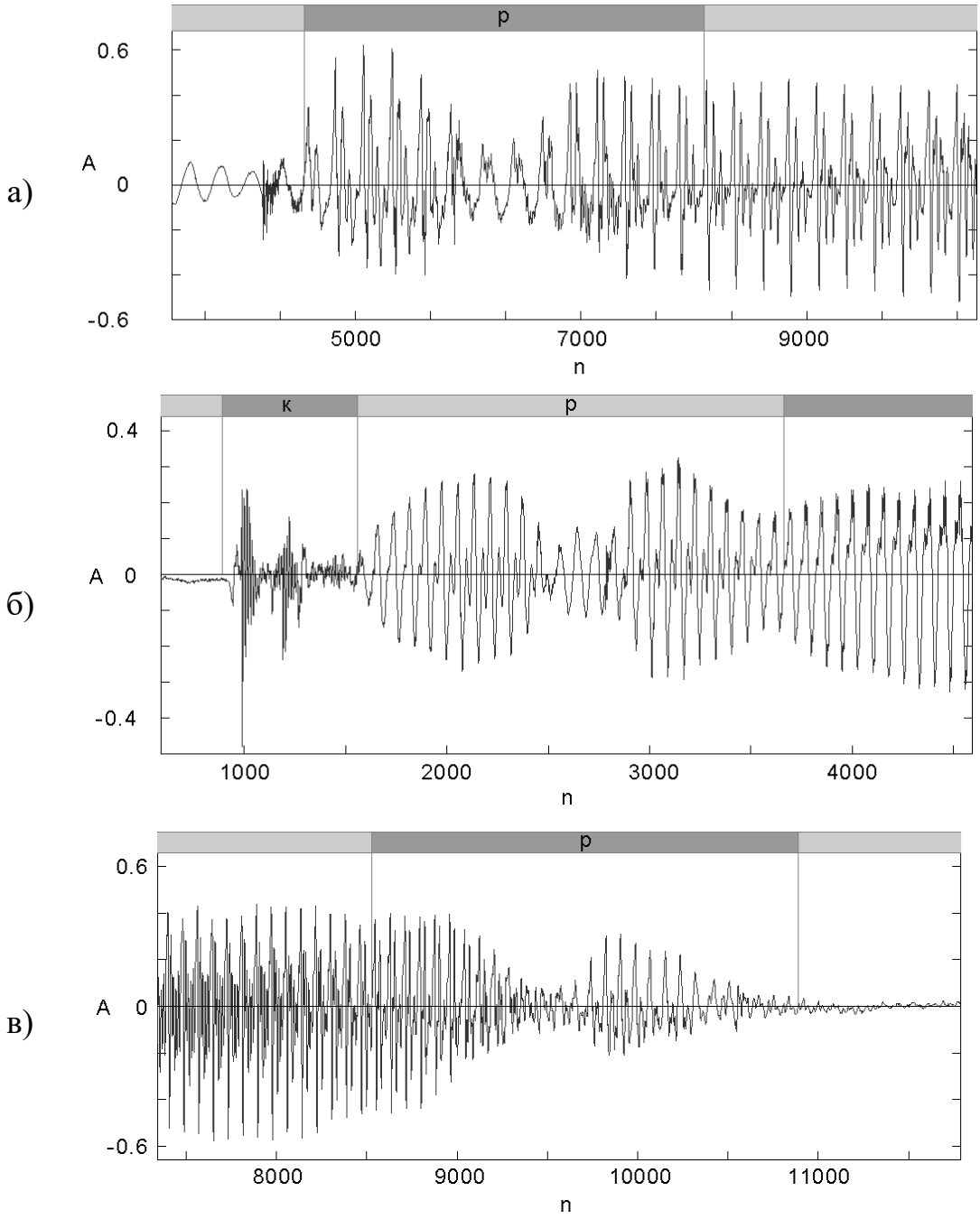


Рисунок 3.4. Примеры реализаций звука [р] разными дикторами: а) слово «Дрозд»; б) слово «Крыша»; в) слово «Марка»

Поэтому был разработан алгоритм, позволяющий получить огибающую, все значения которой состоят из значений исходного сигнала. Такой подход дает больше возможностей для применения выделенной огибающей. В частности, он позволяет оценить период ОТ диктора, осуществить сегментацию на периоды ОТ по максимумам (см. ниже пункт 3.6.2 «Разработка алгоритма ОТ-сегментации »).

При работе алгоритма из отсчетов исходного РС выделяются определенные отсчеты, составляющие в итоге искомую амплитудную МФ. Для уменьшения вычислительной сложности алгоритма выделения огибающей на первом шаге производится выделение множества *M* локальных максимумов РС:

$$M = \{i \mid x(i-1) \le x(i) \wedge x(i+1) < x(i)\}, \tag{3.5}$$

где i – номера отсчетов исходного РС, $x(i)$ – значения отсчетов исходного РС. Далее номера отсчетов огибающей итеративно выделяются из элементов множества *M*. Если принять текущий номер отсчета огибающей за a_n, то номер a_{n+1} следующего отсчета огибающей определяется из условий:

$$\begin{cases} K = \{m \in M \mid a_n < m \le a_n + L\} \\ a_{n+1} \in K \\ \forall k \in K \;\; p(a_{n+1}) \ge p(k) \\ \{k \in K \mid p(a_{n+1}) = p(k) \wedge a_{n+1} < k\} = \varnothing \end{cases}, \tag{3.6}$$

где $p(k)$ определяет наклон линии через точки $(a_n, x(a_n))$ и $(k, x(k))$ (рисунок 3.5):

$$p(k) = \frac{x(k) - x(a_n)}{k - a_n}. \tag{3.7}$$

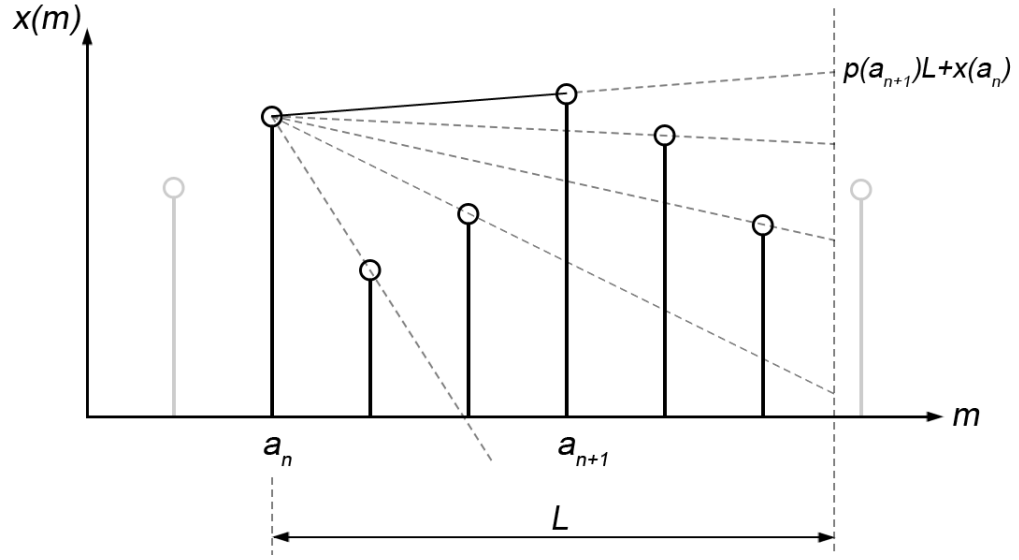


Рисунок 3.5 – К вычислению значений огибающей РС по формулам (3.5-3.7)

В целом, предлагаемый алгоритм, показанный в виде функциональной схемы на рисунке 3.6, представляет собой двухэтапное отсеивание отсчетов сигнала, такое, что остающиеся в результате отсчеты образуют огибающую РС. При этом параметром L регулируется расстояние в отсчетах между точками огибающей, которое алгоритм стремится получить, не превышая его.

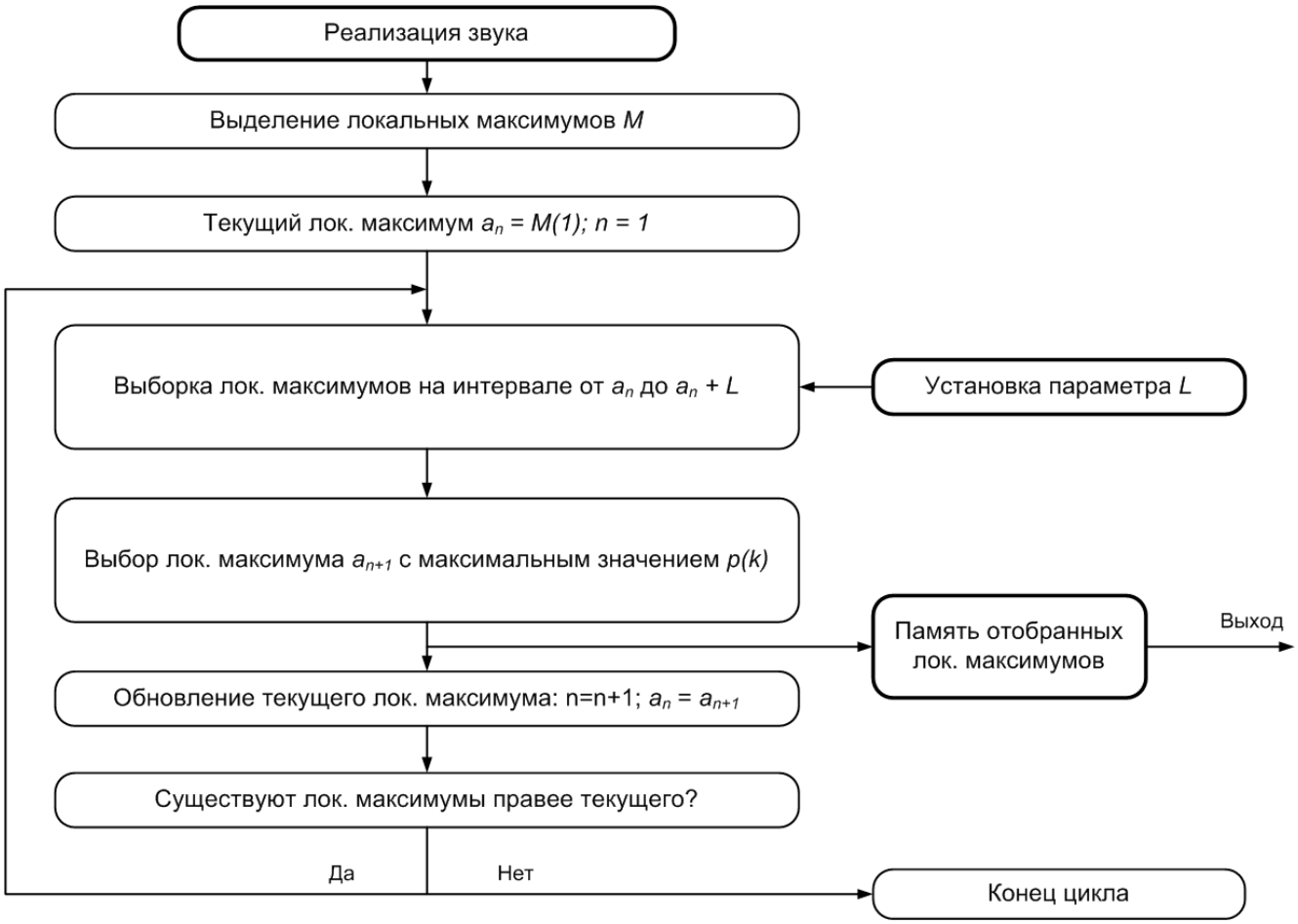


Рисунок 3.6 – Алгоритм выделения огибающей речевого сигнала

На рисунке 3.7 показаны огибающие, полученные и по локальным максимумам, и по локальным минимумам фонограммы слова «*Дал*» при разных соотношениях параметра L и периода ОТ диктора. В случае, когда значение L меньше периода ОТ (на рисунке 3.7а $L = 0{,}5OT$) огибающая не соединяет вершины периодов ОТ вокализованных звуков, а проходит также через локальные максимумы в пределах длительности периодов ОТ. В случае равенства параметра L и средней на сигнале длительности периода ОТ (рисунок 3.7б) огибающая то

соединяет вершины смежных периодов ОТ, когда порог больше их длительностей, то вновь попадает на внутрипериодные локальные максимумы, приобретая пилообразный характер. На рисунке 3.7в показано наиболее полезное на практике соотношение порога и длительности периода ОТ – в данном случае огибающая проходит по вершинам отдельных периодов основного тона. Наконец, если значение порога приближается или превышает удвоенную длительность периода ОТ (на рисунке 3.7 этот случай не отображен), огибающая становится более плавной с пропусками отдельных периодов ОТ.

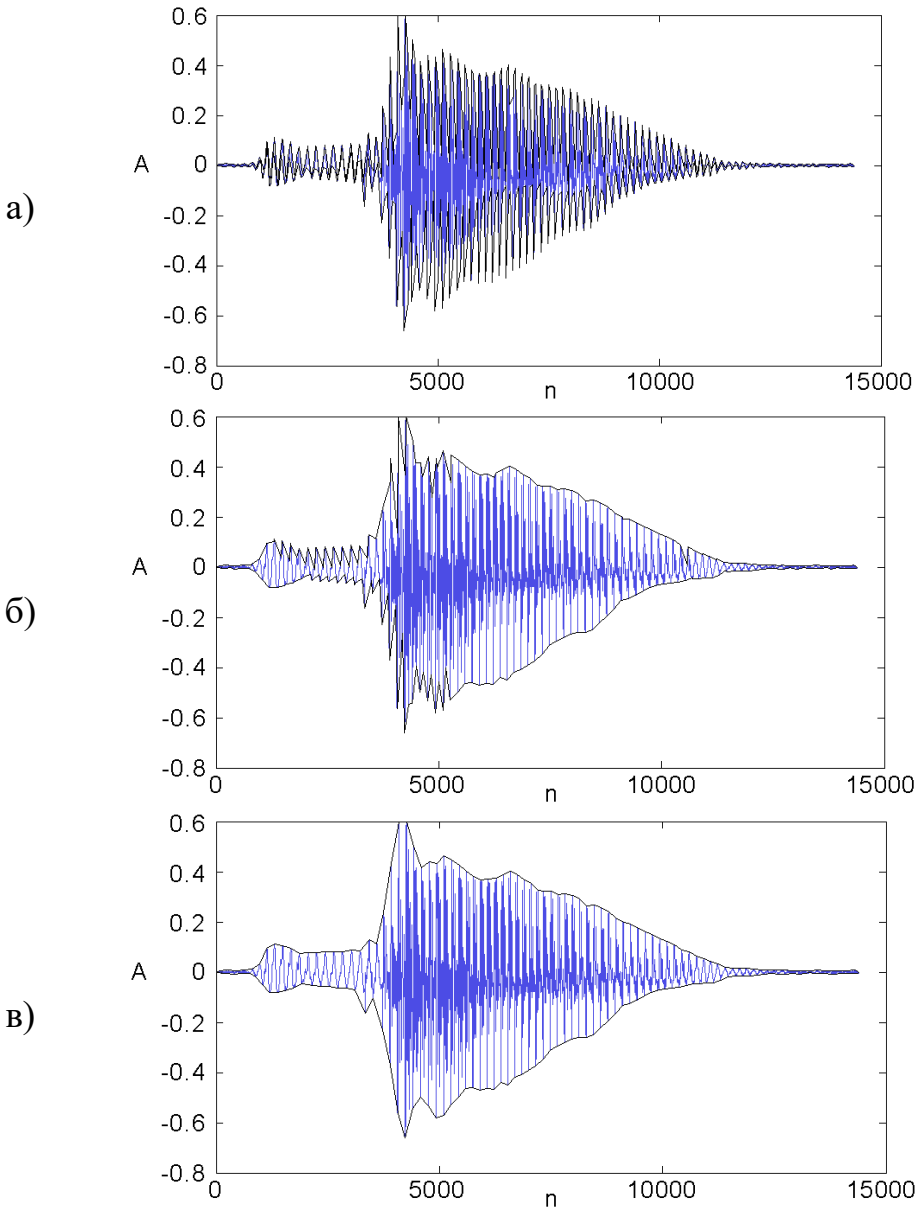


Рисунок 3.7. Результат выделения огибающей на примере слова «Дал»:
а) *L = 0,5OT*; б) *L = OT*; в) *L = 1,5OT*

3.2.2 Применение огибающей в выявлении переходных участков фонограммы

При анализе результатов пофонемной сегментации видно, что большинство переходов от звука к звуку сопровождается изменением амплитудной огибающей – наблюдаются амплитудные разладки. Вариантом выявления таких разладок может быть совместное рассмотрение двух огибающих сигнала, полученных при разных значениях параметра *L*. В таком случае в областях «плавного» изменения амплитудной модуляции сигнала огибающие не будут значимо расходиться друг от друга, а в случае более резкого изменения амплитудной модуляции, огибающие будут вести себя по-разному. На рисунке 3.8 на верхнем графике отображена временная функция речевого сигнала для слова «Несла» и две огибающие: при значении параметров *L=2,5OT* и *L=7,5OT*. На нижнем графике тонкими линиями отображены те же огибающие, а жирными – функция разности огибающих (раздельно для положительных и отрицательных огибающих).

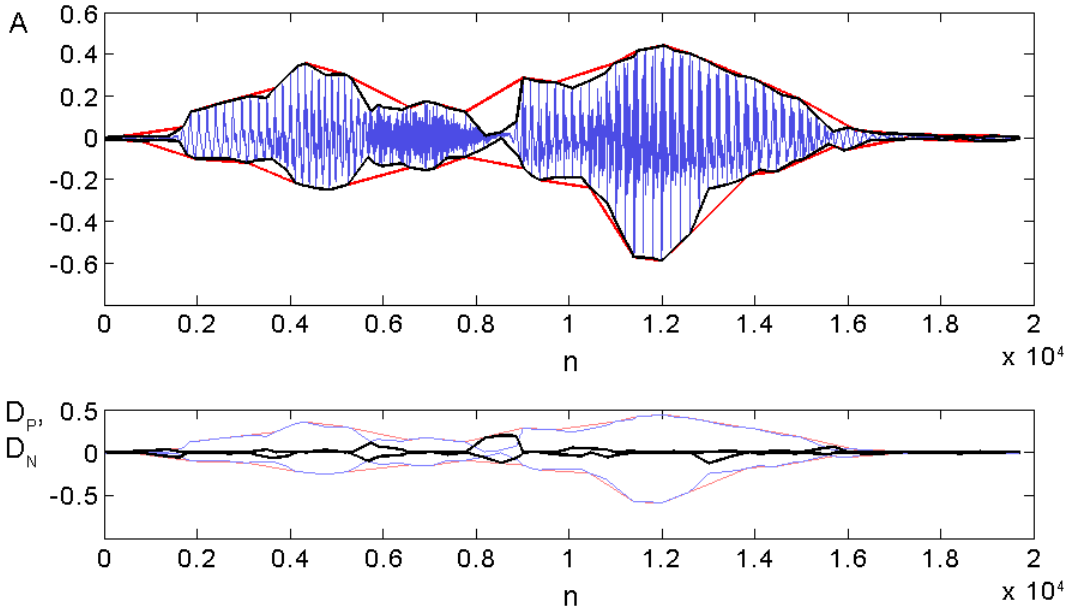


Рисунок 3.8 – Промежуточный этап сегментации по огибающим сигнала

Далее алгоритмом обнуляются отсчеты функции разности положительных огибающих, которым соответствуют нулевые отсчеты функции разности отрицательных огибающих; и наоборот:

$$D_{PF}(n) = \begin{cases} D_P(n), & D_N(n) \neq 0 \\ 0, & D_N(n) = 0 \end{cases};$$
$$D_{NF}(n) = \begin{cases} D_N(n), & D_P(n) \neq 0 \\ 0, & D_P(n) = 0 \end{cases}, \quad (3.8)$$

где n – номер отсчета речевого сигнала, $D_P(n)$ и $D_N(n)$ – функции разности положительных и отрицательных огибающих соответственно.

Наконец, вычисляется разность $D_{PF}(n)$ и $D_{NF}(n)$ (сумма модулей), и по полученной функции производится временная сегментация РС, пример результата которой для фонограммы слова «Несла» показан на рисунке 3.9. Как видно, алгоритм для большей части звуков точно определил границы (показаны в положительной полуобласти). При определенном подходе некорректным можно считать определение правой границы звука [н’], так как звук мягкий и шумовая составляющая этого звука, определяющая мягкость, непосредственно сливается с началом гласного звука [и^э]. При ручной сегментации по этой причине общий для звуков [н’] и [и^э] временной фрагмент был отнесен к звуку [н’], как несущий большую смысловую нагрузку именно для этого звука. Так же стоит отметить ложный маркер сегментации в центральной части ударного звука [а] и маркер, отсекающий звук [а] от не несущих смысловой нагрузки свободных колебаний голосовых связок. Последнее зачастую является положительным фактором.

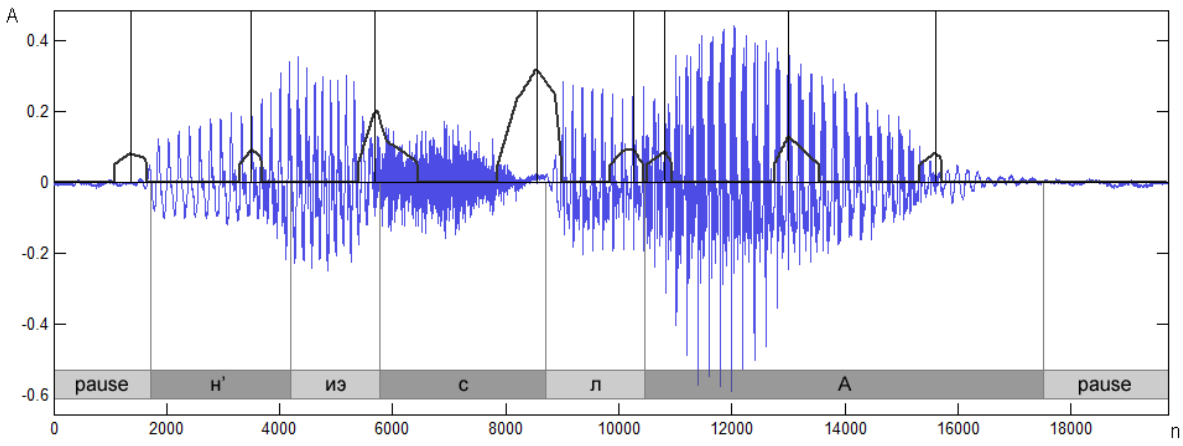


Рисунок 3.9 – Результат сегментации: РС; в положительной полуобласти – функция ($|D_{PF}(n)| + |D_{NF}(n)|$) и метки автоматической сегментации; в отрицательной полуобласти – метки ручной сегментации фонограммы

Важно заметить, что данный алгоритм не может применяться как самодостаточный алгоритм сегментации, так как имеет большие вероятности как пропусков границ звуков в случаях трендового плавного перехода от звука к звуку или малой амплитуды таких звуков как [т’], [к], так и получения ложных маркеров границ звуков в случае сложных эволюций речевого тракта в пределах одного звука. Данный алгоритм может использоваться как вспомогательный для других алгоритмов, учитывающих структуру сигнала, его спектральные или иные характеристики. В частности, данный алгоритм можно применять для коррекции полученных другими алгоритмами маркеров сегментации их привязкой к моментам резкого изменения огибающей РС.

3.3 Повышение результативности использования MFCC-коэффициентов

Как ранее говорилось, одним из наиболее широко распространенных способов параметризации РС является извлечение вектора MFCC-коэффициентов. В существующих речевых приложениях эффективность применения MFCC-параметризации значительно снижается при наличии в РС шумовой составляющей. Вопросу ослабления данной зависимости посвящен ряд исследований, предполагающих применение различных модификаций относительно традиционного алгоритма вычисления MFCC-коэффициентов. К таким модификациям относятся [114]:

- вычисление функции групповой задержки для учета фазовой составляющей спектра РС;
- выполнение преобразования Фурье не по РС, а по его предварительно обработанной автокорреляционной функции, что позволяет уменьшить влияние шумов, убрав из рассмотрения значения автокорреляционной функции на сдвигах менее 3 мс;
- манипуляции формой и количеством фильтров, применяемых в мел-частотной области; введение постобработки результатов фильтрации;

- нормализация длины речевого тракта – одного из факторов, влияющих на дикторозависимость параметров РС;
- внедрение психоакустических моделей (ПАМ), учитывающих известные особенности восприятия человеком акустических сигналов;
- замена кепстрального преобразования вейвлет-преобразованием;
- модификация операции логарифмирования с целью уменьшения влияния шумовой составляющей.

Видно, что спектр возможных модификаций охватывает все шаги традиционного алгоритма вычисления MFCC-коэффициентов. Модификации заключаются либо в незначительных изменениях, либо полных заменах стандартных шагов иными преобразованиями, либо во внедрении в стандартный алгоритм дополнительных промежуточных шагов.

Существующие методы увеличения эффективности MFCC-параметризации показывают некоторое улучшение показателей речевого приложения при наличии в РС шумов. Что характерно, в случае чистого РС традиционный алгоритм MFCC-параметризации показывает лучшие результаты, нежели подавляющее большинство существующих модификаций.

В рамках диссертационной работы предложены две гипотезы, направленные на увеличение результативности применения MFCC-параметризации при обработке зашумленных РС: использование ПАМ звукового анализатора человека, а также воздействие на спектр сигнала на частотах высших гармоник основного тона. Появление второй гипотезы связано с тем, что по длительности большую часть РС русской речи составляют вокализованные звуки, и в то же время важнейшую роль в восприятии звука играют резонансные частоты речевого аппарата (форманты), которые, в свою очередь, влияют на амплитудную модуляцию гармоник ОТ.

3.3.1 Слуховая маскировка и гармоники ОТ

При традиционном вычислении MFCC-коэффициентов используется одна из особенностей восприятия звуков человеком: в силу природы строения

базиллярной мембраны уха человек воспринимает спектры звуков несколькими нелинейно возрастающими по частоте участками – критическими полосами. Однако одна эта особенность лишь отчасти использует знания о психоакустике человеческого слухового анализатора человека.

Как было сказано выше, одним их методов модификации алгоритма MFCC-параметризации является внедрение психоакустической модели эффекта слуховой маскировки. Для этого используется механизм латерального торможения, эффективно описывающий природу одновременного (частотного) маскирования, и реализуемый путем фильтрации спектра мощности РС [114, 115].

Однако также известен другой подход к учету одновременного маскирования, широко применяемый в системах сжатия аудиосигналов, и заключающийся в вычислении глобального маскирующего порога. Работа данной ПАМ основана на том факте, что человек не слышит одновременно весь диапазон частот, так как происходит частотное маскирование сигнала: при одновременном присутствии двух сигнальных составляющих на близких частотах более слабый сигнал становится неслышимым на фоне более сильного.

Было принято решение рассмотреть данную ПАМ как альтернативу реализации механизма латерального торможения в модификациях алгоритма MFCC-параметризации. Для этих целей произведена адаптация широко известного алгоритма, применяемого в стандарте сжатия аудиосигналов ISO/IEC MPEG-1 Layer 1 [116], математическое описание работы данного алгоритма приведено в статьях [117, 118].

Для сравнения с эффективностью внедрения механизма латерального торможения рассматривается подалгоритм LI (Lateral Inhibition, латеральное торможение), входящий в состав алгоритма LTFC – модифицированного алгоритма MFCC-параметризации [119].

Наконец, как было сказано выше, второй гипотезой увеличения эффективности применения MFCC-параметризации является воздействие на спектр сигнала на частотах гармоник ОТ для получения более выраженной картины формантных частот. Для воздействия на значения спектра мощности на

частотах, кратных частоте ОТ, предложено следующее преобразование спектральной плотности мощности РС:

$$P(k)=\begin{cases}(1+0,5|i|)P_x(k)Kp_v, & k=\left[nN\frac{f_{OT}}{F_S}\right]+i, n\in\mathbb{N}, n<\frac{F_S}{2f_{OT}}, i=\overline{-1,1}\\ P_x(k) & k\neq\left[nN\frac{f_{OT}}{F_S}\right]+i, n\in\mathbb{N}, n<\frac{F_S}{2f_{OT}}, i=\overline{-1,1}\end{cases}, \quad (3.9)$$

где $P_x(k)$ – k-ый отсчет спектральной плотности мощности РС для текущего временного окна; N – длина окна Фурье-преобразования; F_s – частота дискретизации РС; f_{OT} – оценка частоты ОТ в текущем временном окне, $0 \le p_v \le 1$ – метрика наличия вокализации РС в текущем временном окне (0 для невокализованного фрагмента, 1 для вокализованного); $K \ge 0$ – коэффициент преобразования, эмпирически был подобрано значение 1,3; квадратными скобками показана операция округления до ближайшего целого.

3.3.2 Экспериментальное исследование

Для исследования результатов модификации алгоритма вычисления MFCC-коэффициентов использована система распознавания раздельно произнесенных слов из ограниченного словаря, написанная на языке MATLAB доцентом Ли-Мин Ли (Lee-Min Lee) из Da-Yeh University, Тайвань [120, 121]. В силу необходимости использования в данном исследовании достаточно большой базы речевых фонограмм, принято решение воспользоваться англоязычной базой TIDIGITS, содержащей группу из 2072 тренировочных и 2486 тестовых фонограмм с частотой дискретизации 8 кГц. Каждая фонограмма базы содержит одно слово из словаря, который включает в себя 11 слов: цифры от нуля до девяти (ноль при этом произносится в двух вариантах: «oh» и «zero»). При составлении базы использованы голоса 94 мужчин и 114 женщин, причем дикторы тренировочной и тестовой групп не пересекаются.

Для моделирования шумовой обстановки в тестовые фонограммы добавлялись шумы различной природы: уличный шум, шум в поезде, шум в автомобиле, шум толпы (множественные фоновые голоса) – для ОСШ от 20 дБ до 0 дБ с шагом 5 дБ. Представленные ниже результаты являются усредненными по перечисленным шумам, в свою очередь, полные неусредненные таблицы приведены в Приложении В. Обучение системы распознавания производилось на исходных чистых тренировочных фонограммах.

Программный код использованной системы распознавания изменен таким образом, что при параметризации РС используются только MFCC-коэффициенты.

Для оценки значения частоты ОТ фрагмента РС при исследовании гипотезы использован алгоритм PEFAC, описанный в работе [122] и реализованный в тулбоксе VoiceBox вычислительной среды MATLAB. Указанный алгоритм не только определяет оценку частоты ОТ фрагмента РС, но также возвращает метрику p_v принадлежности фрагмента к вокализованным звукам. Тем не менее, для решения данной задачи могут быть опробованы и иные алгоритмы оценки текущей частоты ОТ, удовлетворяющие приведенному условию.

В таблице 3.1 показаны пословные точности распознавания различными алгоритмами: MFCC(13) – традиционный алгоритм вычисления тринадцати MFCC-коэффициентов; LI – внедрение алгоритма латерального торможения [119]; MPEG1 – внедрение ПАМ модели стандарта ISO/IEC MPEG-1 Layer 1 [116]; FFH – внедрение алгоритма усиления гармоник основного тона (Fundamental Frequency Harmonics), формула (3.9); LI+FFH и MPEG1+FFH – совместное использование алгоритма FFH соответственно с алгоритмами LI и MPEG1.

В таблице 3.2 представлены значения относительного улучшения результатов распознавания рассматриваемыми модифицированными алгоритмами по сравнению с алгоритмом MFCC(13). Здесь и далее относительное улучшение *RI* рассчитывается по формуле:

$$RI = \frac{RR_A - RR_{MFCC}}{100 - RR_{MFCC}} \times 100\%, \qquad (3.10)$$

где RR_A – точность распознавания, полученная для рассматриваемого модифицированного алгоритма и выраженная в процентах; RR_{MFCC} – точность распознавания в процентах, полученная алгоритмом MFCC при тех же условиях.

Таблица 3.1. Пословная точность распознавания (%) при различных ОСШ

Алгоритм\ОСШ	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	средн. 0-20 дБ
MFCC(13)	90,7	75,7	68,4	58,4	44,6	31,2	55,7
LI	89,6	75,0	68,4	59,4	46,6	33,1	56,5
MPEG1	84,5	75,8	71,9	65,1	55,0	40,3	61,6
FFH	90,4	75,8	68,9	59,0	46,0	32,5	56,4
LI+FFH	89,1	75,1	68,3	59,7	47,3	34,0	56,9
MPEG1+FFH	84,6	75,7	71,9	65,1	55,3	40,8	61,8

Таблица 3.2. Относительные улучшения (%) в сравнении с алгоритмом MFCC(13)

Алгоритм\ОСШ	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	средн. 0-20 дБ
LI	-12,6	-2,7	0,1	2,3	3,6	2,6	1,2
MPEG1	-67,8	0,3	11,2	16,0	18,9	13,2	11,9
FFH	-3,9	0,3	1,8	1,3	2,6	1,9	1,6
LI+FFH	-17,8	-2,4	-0,2	2,9	4,9	4,0	1,8
MPEG1+FFH	-66,5	0,1	11,2	16,1	19,4	13,8	12,1

Традиционный алгоритм MFCC-параметризации показывает лучшие результаты для чистого РС, данное положение подтверждается и другими исследованиями [114, 123]. Однако эффективность использования традиционного алгоритма стремительно падает при уменьшении ОСШ. В этом случае лучших результатов можно добиться, используя модифицированные алгоритмы. Как видно из таблиц 3.1 и 3.2, внедрение в алгоритм ПАМ, описанной в стандарте ISO/IEC MPEG-1 Layer 1, позволяет достигнуть значимого улучшения работы речевого приложения в шумовом окружении.

Предложенное преобразование спектра мощности РС, заключающееся в усилении спектральных составляющих на частотах, кратных частоте ОТ, также позволяет повысить результаты речевого приложения, и, как видно из

представленных таблиц, данное преобразование при низких ОСШ может быть использовано совместно с психоакустическими модификациями LI и MPEG1: в таблицах 3.1 и 3.2 строки LI+FFH, MPEG1+FFH.

Так как используемый в алгоритме FFH алгоритм оценки частоты ОТ также подвержен влиянию фоновых шумов, для оценки потенциального эффекта от использования FFH был рассмотрен идеализированный алгоритм оценки частоты ОТ – соответствующая вариация обозначается далее FFHi. В данной идеализации на вход используемого алгоритма оценки частоты ОТ подается соответствующий фрагмент чистой фонограммы, а все остальные блоки по-прежнему работают с фрагментами с заданным ОСШ. Результаты представлены в таблицах 3.3 и 3.4.

Таблица 3.3. Пословная точность распознавания (%) при оценке частоты ОТ на чистой фонограмме

Алгоритм\ОСШ	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	средн. 0-20 дБ
FFHi	90,4	76,2	69,6	59,8	46,9	33,5	57,2
LI+FFHi	89,1	75,3	68,7	60,2	47,9	34,8	57,4
MPEG1+FFHi	84,6	75,8	72,1	65,4	55,5	41,4	62,0

Таблица 3.4. Относительные улучшения (%) при оценке частоты ОТ на чистой фонограмме

Алгоритм\ОСШ	Чистый	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	средн. 0-20 дБ
FFHi	-3,9	1,9	4,0	3,4	4,3	3,3	3,4
LI+FFHi	-17,8	-1,8	1,1	4,3	6,0	5,1	3,0
MPEG1+FFHi	-66,5	0,2	11,7	16,6	19,8	14,8	12,6

Производительность алгоритмов оценивалась по суммарному времени, затрачиваемому вычислительной машиной на параметризацию всей базы тренировочных и тестовых фонограмм. Полученные значения, нормированные к значению для традиционного алгоритма MFCC(13), приведены в таблице 3.5.

Таблица 3.5. Относительное время работы алгоритмов MFCC-параметризации

Алгоритм	MFCC(13)	LI	MPEG1	FFH	LI+FFH	MPEG1+FFH
Время	1,0	1,1	7,6	3,8	3,9	11,4

ПАМ, описываемая стандартом ISO/IEC MPEG-1 Layer 1, требует значительно больше вычислительных затрат, нежели фильтрация спектра мощности, реализующая алгоритм LI. Тем не менее, реализации данной ПАМ обладают достаточным быстродействием для применения в системах реального времени [124]. Быстродействие модификации FFH, в свою очередь, непосредственно определяется быстродействием применяемого в ней алгоритма оценки частоты ОТ.

Для дополнительного увеличения эффективности MFCC-параметризации могут также использоваться методы, рассматривающие РС вне пределов одного текущего фрейма. В частности, известен положительный результат от применения механизма ПАМ временного (неодновременного) маскирования [119, 123].

3.4 Сегментация первого уровня – определение границ речевой активности

Задача выделения границ речевой активности состоит в определении в речевом сигнале моментов времени, соответствующих началам и завершениям фрагментов сигнала с присутствующей речевой информационной составляющей.

3.4.1 Сложности реализации

Для реализации задачи определения границ речевой активности алгоритм VAD должен определять наличие сигнальных признаков, характерных только для активной речи (или наоборот, только для пауз). Одним из самых простых таких признаков является энергетика сигнала. В идеальном случае на участках пауз мощность сигнала равна нулю. Однако наличие фоновых шумов осложняет применение энергии сигнала в качестве параметра для VAD-сегментации и является причиной применения более сложных алгоритмов. Пример сильно зашумленной (городской шум при записи внутри помещения) фонограммы приведен на рисунке 3.10. В фонограмме с паузами между словами произносится 12 цифр. На амплитудно-временной диаграмме достаточно сложно выделить участки активной речи и пауз, однако при прослушивании фонограммы цифры

определяются четко. При рассмотрении сигнала в частотно-временной области (рисунок 3.11) слова отчетливо выделяются на фоне шумовой составляющей.

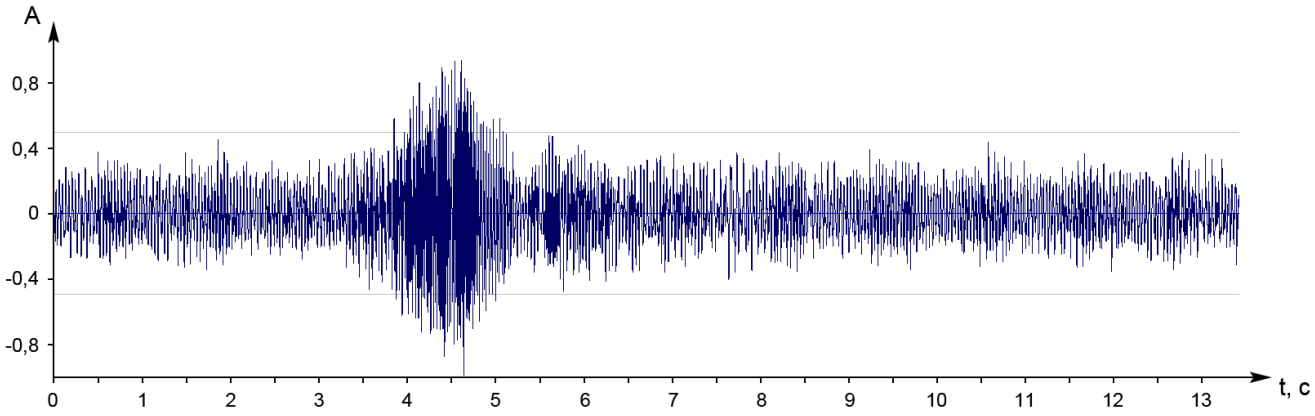


Рисунок 3.10 – Фрагмент зашумленного РС, содержащий 12 слов

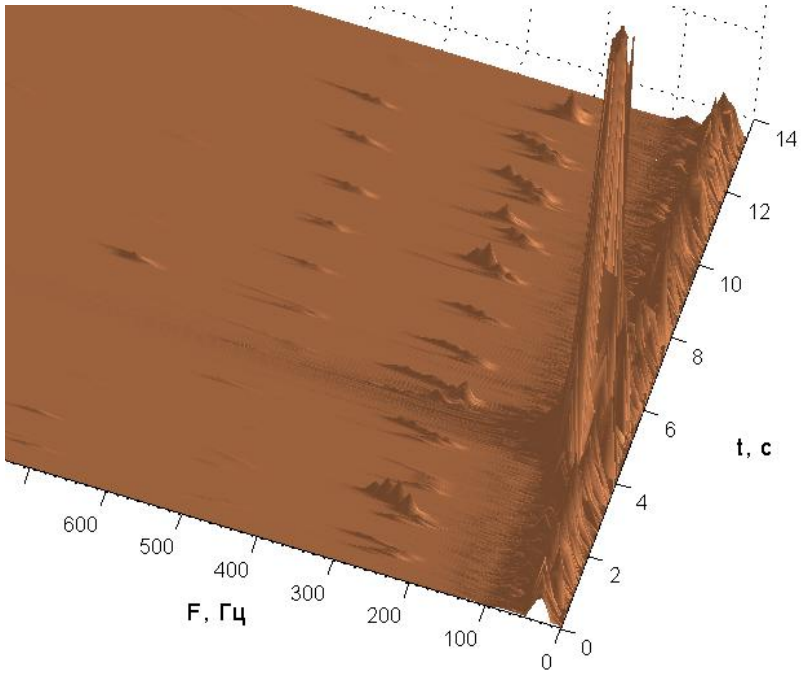


Рисунок 3.11 – Срез трехмерной спектрограммы зашумленного фрагмента РС, представленного на рисунке 3.10

В общем случае, усложнять состоятельную VAD-сегментацию фонограмм может наличие нескольких факторов, среди которых:

– фоновые акустические шумы;
– плавное нарастание мощности РС в начале речевой активности;
– начало речевой активности с шумного или взрывного звука, интенсивность и/или длительность которого невелика.

Фоновые акустические шумы широко варьируются по причинам своего появления и, соответственно, оказывают различное влияние на фонограммы. Рассмотрение речевых фонограмм, полученных при различных условиях записи и различных характеристиках записывающей и передающей аппаратуры, позволяет составить перечень основных типов шумов в фонограммах: таблица 3.6.

Таблица 3.6. Шумы в фонограмме

Шумы аппаратуры и канала связи	1.	Треск: множественные кратковременные импульсные помехи
	2.	Длительное «жужжание»
	3.	Кратковременные шумы, воспринимаемые как «позвякивания»
	4.	Непрерывные наводки от аппаратуры
	5.	Превышение уровнем сигнала допустимого динамического диапазона
Внешние шумы	1.	Фоновая речь посторонних дикторов
	2.	Шумовые щелчки (имеют шумовое заполнение)
	3.	Городской шум (автотранспорт, движение пешеходов)
	4.	Шум природных явлений (дождь, ветер)
	5.	Фоновый офисный шум (музыка, работа техники)
	6.	Шум, вызванный шагами
	7.	Охриплые вдохи и придыхания диктора

3.4.2 Повышение эффективности энергетического VAD-алгоритма

В рамках работ [125] и [126], выполненных на базе государственного университета аэрокосмического приборостроения, представлен энергетический VAD-алгоритм. Функциональная схема алгоритма показана на рисунке 3.12, программная реализация алгоритма на языке MATLAB приведена в приложениях в [126].

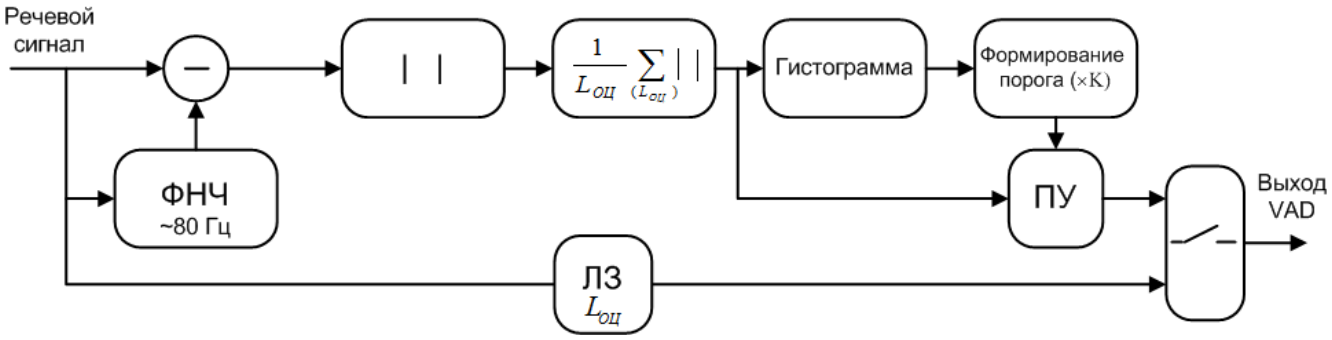


Рисунок 3.12 – Функциональная схема энергетического алгоритма VAD

В алгоритме речевой сигнал проходит предварительную ВЧ фильтрацию, необходимую для подавления низкочастотных наводок (операция осуществляется ФНЧ и вычитателем).

Далее на интервалах оценивания текущих средних значений отсчетов (длина интервала $L_{ОЦ}$ выбирается в пределах от 2,5 до 7 мс в зависимости от частоты дискретизации: по достаточной по шуму выборке в 50...70 отсчетов) осуществляется суммирование модулей отсчетов РС с последующей нормировкой на длину интервала. Гистограмма строится по результатам обработки порядка 1,5-2 секунд фонограммы, в течение которых внешний шум считается стационарным. На рисунке 3.13 показаны по два примера графиков фильтрованного интервалами оценивания модуля РС и их гистограмм. На рисунке 3.13а обрабатываемое слово «Здравствуйте» дискретизировано с частотой 8000 Гц, а на рисунке 3.13б представлены графики для фразы «Один – два», записанной на другой аппаратуре с частотой дискретизации 22050 Гц. На рисунке 3.13 на верхних двух графиках показан фильтрованный интервалами оценивания РС и уровень вычисленного порога. На нижних двух графиках показаны гистограммы, по которым порог вычисляется. Самый простой – но не самый надежный – способ автоматического установления порога по гистограмме – поиск на ней в направлении слева направо (то есть от меньших значений уровня сигнала к большим) первого минимума после первого максимума, как и показано вертикальной линией на рисунке 3.13.

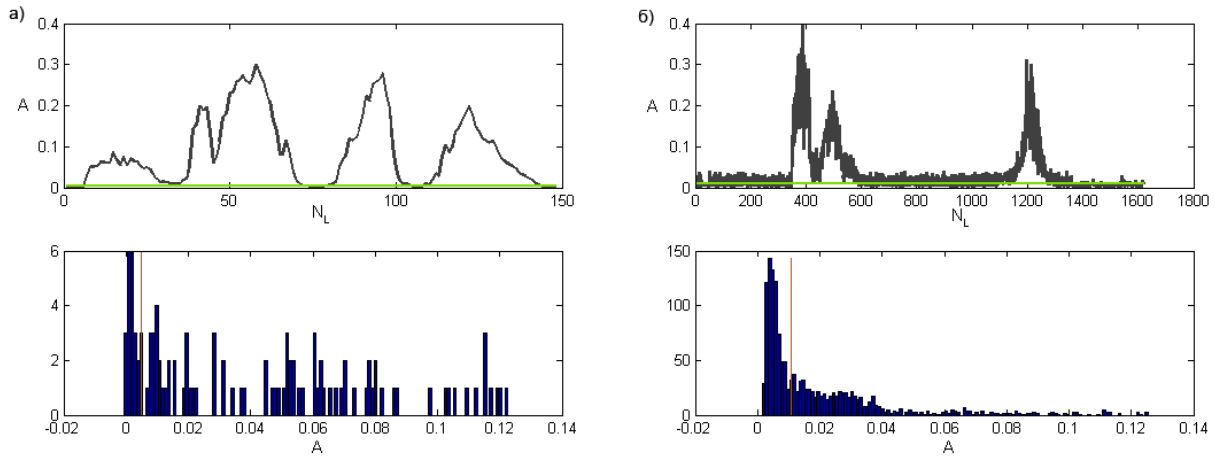


Рисунок 3.13 – Установление порога в VAD: а) на примере слова «Здравствуйте»; б) на примере фразы «Один – два»

Более надежным является обоснование уровня порога непосредственно по уровню максимума гистограммы. При этом для достижения высокой точности определения данного значения необходимо решать дилемму между стремлением сузить ширину дифференциального коридора гистограммы и сохранить достаточно большим количество точек, попадающих в эти коридоры в области максимума (например, на рисунке 3.13а в коридор, приходящийся на максимум гистограммы, приходится всего 6 осредненных значений модуля сигнала).

Для достижения достаточной точности необходимо автоматически подбирать ширину дифференциального коридора гистограммы так, чтобы ее максимум приходился по порядку на третий или четвертый коридор. Порог устанавливается равным значению уровня сигнала, соответствующему максимуму гистограммы и домноженному на постоянный коэффициент, устанавливаемый в пределах от 1,7 до 2,3.

Для получения достаточного количества точек, попадающих в дифференциальный коридор гистограммы, необходимо использовать в фонограммах сравнимые по длительности с активной речью участки пауз (это условие выполняется в естественной речи).

Описанный энергетический VAD-алгоритм показывает хорошие результаты при выделении слов, начинающихся и заканчивающихся большими по амплитуде реализациями звуков (в основном, вокализованными). Однако, в случае наличия

шумных звуков, сравнимых с уровнем фонового шума, резко возрастает вероятность пропуска таких звуков. Кроме того, результаты сегментации резко ухудшаются по мере увеличения фонового шума.

Перечисленные проблемы в значительной степени позволяет решить модифицированный энергетический VAD-алгоритм, функциональная схема которого приведена на рисунке 3.14.

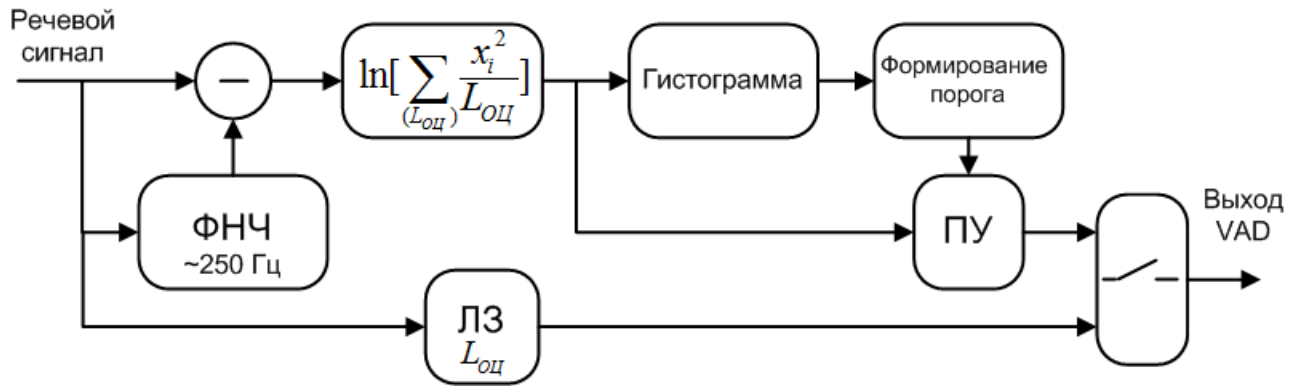


Рисунок 3.14 – Функциональная схема модифицированного алгоритма VAD

В данном модификации, во-первых, значительно повышена частота среза НЧ-фильтра: с частоты среза, лежащей в пределах от 60 до 80 Гц, до значения 250 Гц. Несмотря на то, что частота ОТ в подавляющем большинстве случаев оказывается в этом случае подавленной, алгоритм остается работоспособным за счет присутствия высших гармоник (см. выше раздел «1.1.3 Произнесение и восприятие речи человеком. Фонетическое строение сигнала русской речи»). Как будет показано ниже, такое решение значительно повышает устойчивость алгоритма к встречающимся на практике фоновым шумам.

Далее имеющийся фрагмент сигнала по-прежнему разбивается прямоугольными неперекрывающимися окнами оценивания длительностью w = 9 мс, на каждом из которых вычисляется логарифм энергии сигнала:

$$E_w = \ln\left(\sum_{i=k}^{k+w} x_i^2\right), \tag{3.11}$$

где k – позиция начала текущего интервала оценивания, x_i – отсчеты РС.

Операция логарифмирования согласуется с механизмом восприятия интенсивности звука человеком (эмпирический психофизиологический закон Вебера-Фехнера) и позволяет увеличить разрешающую способность гистограммы в области небольших значений энергий, в которой, как было сказано выше, и возникают основные причины ошибок работы VAD-алгоритма.

В результате при построении гистограммы большинство значений энергии E_w вокализованных звуков могут попасть в один-два смежных коридора, образуя на гистограмме абсолютный максимум. Чтобы заведомо убрать влияние такого максимума на выбор порога по гистограмме, производится отсечение интервалов оценивания, соответствующих установившимся вокализованным колебаниям.

По результатам наблюдений, в вокализованных звуках содержится приблизительно 95-99,5% энергии РС. Для устранения влияния на выбор порога по гистограмме значений энергий вокализованных интервалов оценивания в текущей реализации алгоритма предлагается не учитывать для построения гистограммы наибольшие значения E_w, составляющие в сумме 90% от общей энергии фрагмента. На рисунке 3.15 показаны вычисленные значения энергий на интервалах оценивания. Красными жирными штрихами отмечены неиспользуемые для построения гистограммы значения E_w.

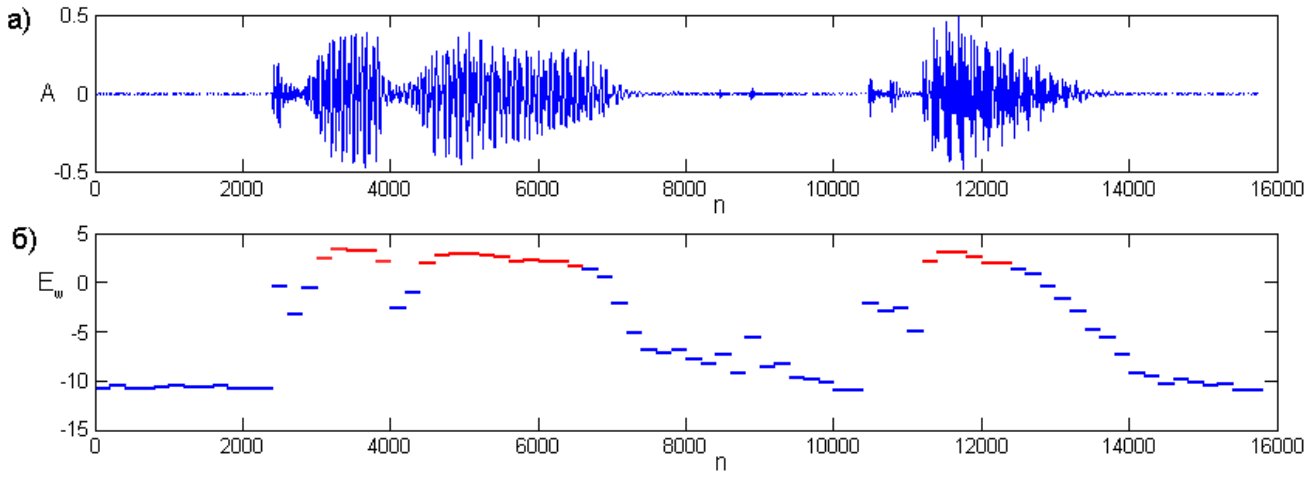


Рисунок 3.15 – Интервалы оценивания в VAD-алгоритме: а) РС, соответствующий слову «Трубка»; б) значения энергии E_w на интервалах оценивания; красным показаны наибольшие значения E_w, составляющие в сумме 90% от общей энергии

Пример гистограммы, построенной для слова «Трубка», приведен на рисунке 3.16. В качестве порога принимается координата центра первого минимального канала гистограммы правее главной моды (вертикальная пунктирная линия на рисунке).

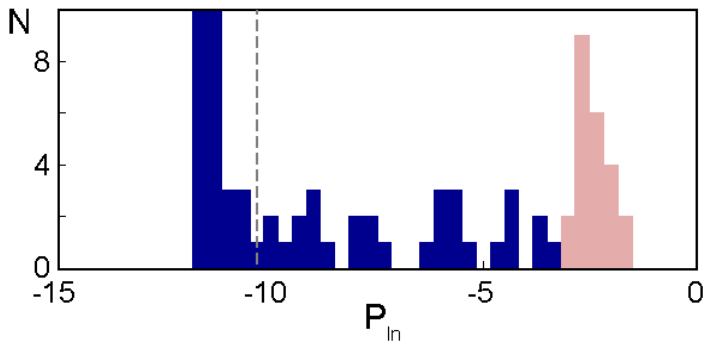


Рисунок 3.16 – Гистограмма значений E_w

В качестве участков речевой активности принимаются отсчеты сигнала, соответствующие интервалам, значение E_w которых превышает вычисленный по гистограмме порог.

В заключение полученные фрагменты речи / пауз подвергаются дополнительному анализу:

– удаляются одиночные интервалы активности;
– удаляются фрагменты пауз длительностью не более двух интервалов оценивания (согласно исследованию, описанному в разделе 2 «Исследование сигнальных особенностей звуков русской речи» минимальная длительность смычки имеет значения порядка 20-25 мс);
– удаляются фрагменты речевой активности длительностью не более трех интервалов оценивания (соответственно, минимальная длительность звука, обрамленного паузами и / или смычками, имеет значения порядка 30-35 мс).

Пример результата автоматической сегментации речевого сигнала, соответствующего слову «Трубка», представлен на рисунке 3.17.

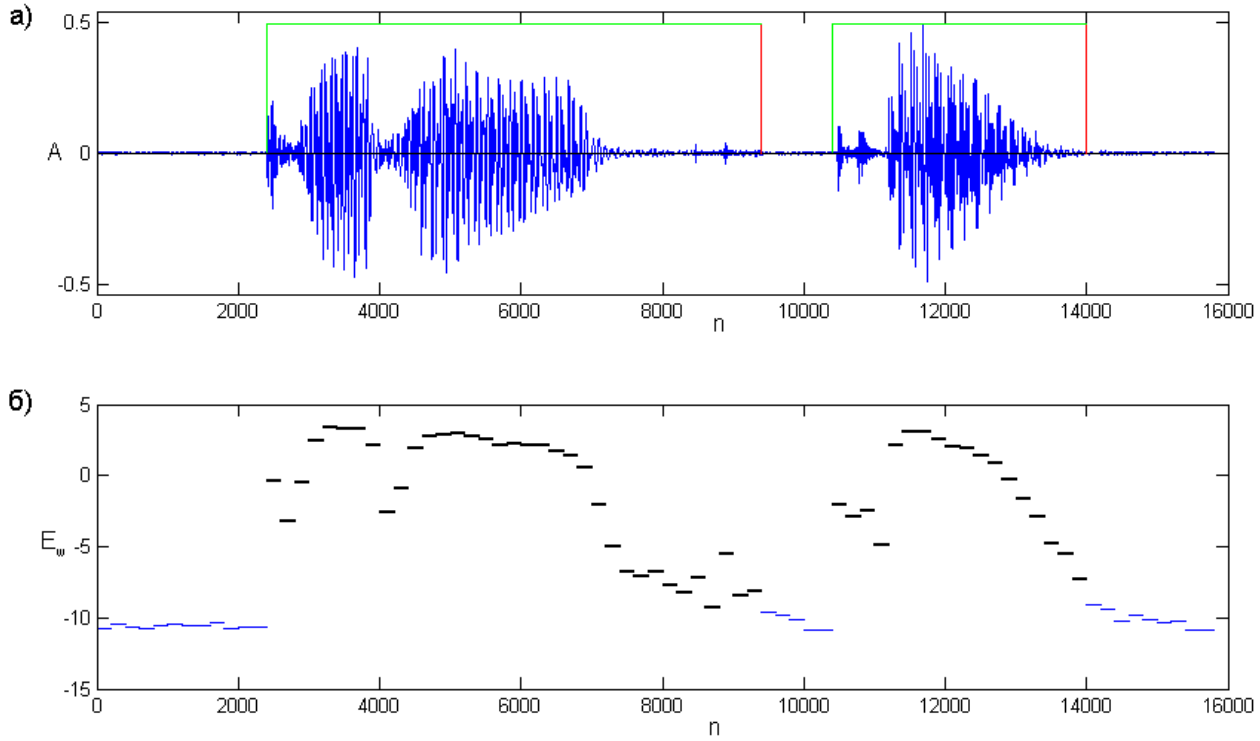


Рисунок 3.17 – Результат автоматической VAD-сегментации слова «Трубка»: а) исходный речевой сигнал и обозначение автоматически определенных границ речевой активности; б) значения энергии E_w на интервалах оценивания; черным жирным показаны интервалы, соответствующие речевой активности

3.4.3 Сравнение эффективности разработанных VAD-алгоритмов

Для сравнения эффективности VAD-алгоритмов с помощью разработанного в рамках диссертации метода (подраздел 3.1.3 «Метод сравнения эффективности работы однотипных алгоритмов сегментации»), вычислены ошибки определения границ начала речевой активности *_VS*, границ окончания речевой активности *VS_* и общая ошибка сегментации *VS* (таблицы 3.7 и 3.8). Анализ произведен на небольшом речевом материале, содержащем трудные для VAD-алгоритмов особенности РС: шумные и короткие слабые взрывные звуки на границах речевой активности, значительный фоновый шум.

Во всех случаях ошибки вычисляются относительно эталонной длительности всех сегментов активной речи – в соответствии с определением VAD-алгоритма как алгоритма выделения активной речи, но не пауз. Стоит

отметить, что ошибка *VS* включает не только ошибки *_VS* и *VS_*, но также учитывает возможные пропуски активных сегментов и ложные активные сегменты. Ошибки определения границ начала речевой активности и ее окончания вычисляются раздельно, так как в общем случае вокализованный звук может заканчиваться небольшими по амплитуде релаксационными колебаниями голосовых связок. При этом выбор правой границы речевой активности становится неоднозначной задачей. Подробнее этот вопрос рассматривается в следующем подразделе 3.4.4 «Ограничение остаточных колебаний вокализованных звуков перед паузой и смычкой».

Таблица 3.7 составлена для сравнения двух вариантов энергетического VAD-алгоритма: исходного, представленного в [125, 126] (VAD-алгоритм A, рисунок 3.12), и модифицированного (VAD-алгоритм B, рисунок 3.14).

В таблице 3.8 показаны результаты измерения ошибок сегментации для VAD-алгоритма B и VAD-алгоритма, используемого в расширении VoiceBox программной среды MATLAB. Реализация алгоритма VAD, используемая в VoceBox, основана на описанной в [127] разработке. В качестве параметров, являющихся основой для процесса сегментации, применяются Мел-кепстральные коэффициенты. В упомянутой работе также приведены данные по сравнению эффективности их алгоритма и стандартизированного VAD-алгоритма G.729B.

Примеры графиков, иллюстрирующих соответствия эталонной (ручной) сегментации и результатов автоматической сегментации для разных алгоритмов приведены на рисунках 3.18...3.20. Стоит отметить, что фонограмма слова «Шесть» отличается от остальных присутствием сравнительно большого фонового шума. Также необходимо подчеркнуть, что:

- результаты ручной сегментации являются субъективными и невоспроизводимыми;
- VAD-алгоритмы могут быть предназначены для работы в определенных специфических условиях, давая в иных условиях худший результат;
- остаточные свободные колебания в конце вокализованных звуков перед

паузой порождают несколько подходов к определению правой границы вокализованного звука, поэтому различия в работе алгоритмов в данном случае не всегда имеет смысл считать ошибкой.

Таблица 3.7. Сравнение эффективности разработанных VAD-алгоритмов

	VAD-алгоритм A				**VAD-алгоритм B**			
Фонограмма	**_VS**	**VS_**	**VS**	**Пропусков + ложных**	**_VS**	**VS_**	**VS**	**Пропусков + ложных**
«Трубка» диктор-мужчина	3,5%	3,6%	10,0%	1	0,4%	8,1%	8,5%	0
«Трубка» диктор-женщина	0,7%	8,2%	12,3%	2	0,3%	13,6%	13,9%	0
«Фетр» диктор-мужчина	2,0%	2,1%	4,1%	0	3,8%	14,0%	17,8%	0
«Шесть» диктор-женщина	33,0%	35,5%	90,3%	4	8,8%	5,8%	14,6%	0
«Фельдъегерь» диктор-женщина	0,2%	0,1%	1,8%	1	0,0%	0,1%	0,1%	0

Таблица 3.8. Сравнение эффективности разработанного VAD-алгоритма B и VAD-алгоритма VoiceBox

	VAD-алгоритм VoiceBox				**VAD-алгоритм B**			
Фонограмма	**_VS**	**VS_**	**VS**	**Пропусков + ложных**	**_VS**	**VS_**	**VS**	**Пропусков + ложных**
«Трубка» диктор-мужчина	10,2%	9,8%	34,6%	2	0,4%	8,1%	8,5%	0
«Трубка» диктор-женщина	2,4%	9,8%	12,1%	0	0,3%	13,6%	13,9%	0
«Фетр» диктор-мужчина	5,4%	11,9%	17,3%	0	3,8%	14,0%	17,8%	0
«Шесть» диктор-женщина	4,6%	3,2%	33,2%	8	8,8%	5,8%	14,6%	0
«Фельдъегерь» диктор-женщина	6,6%	11,9%	18,5%	0	0,0%	0,1%	0,1%	0

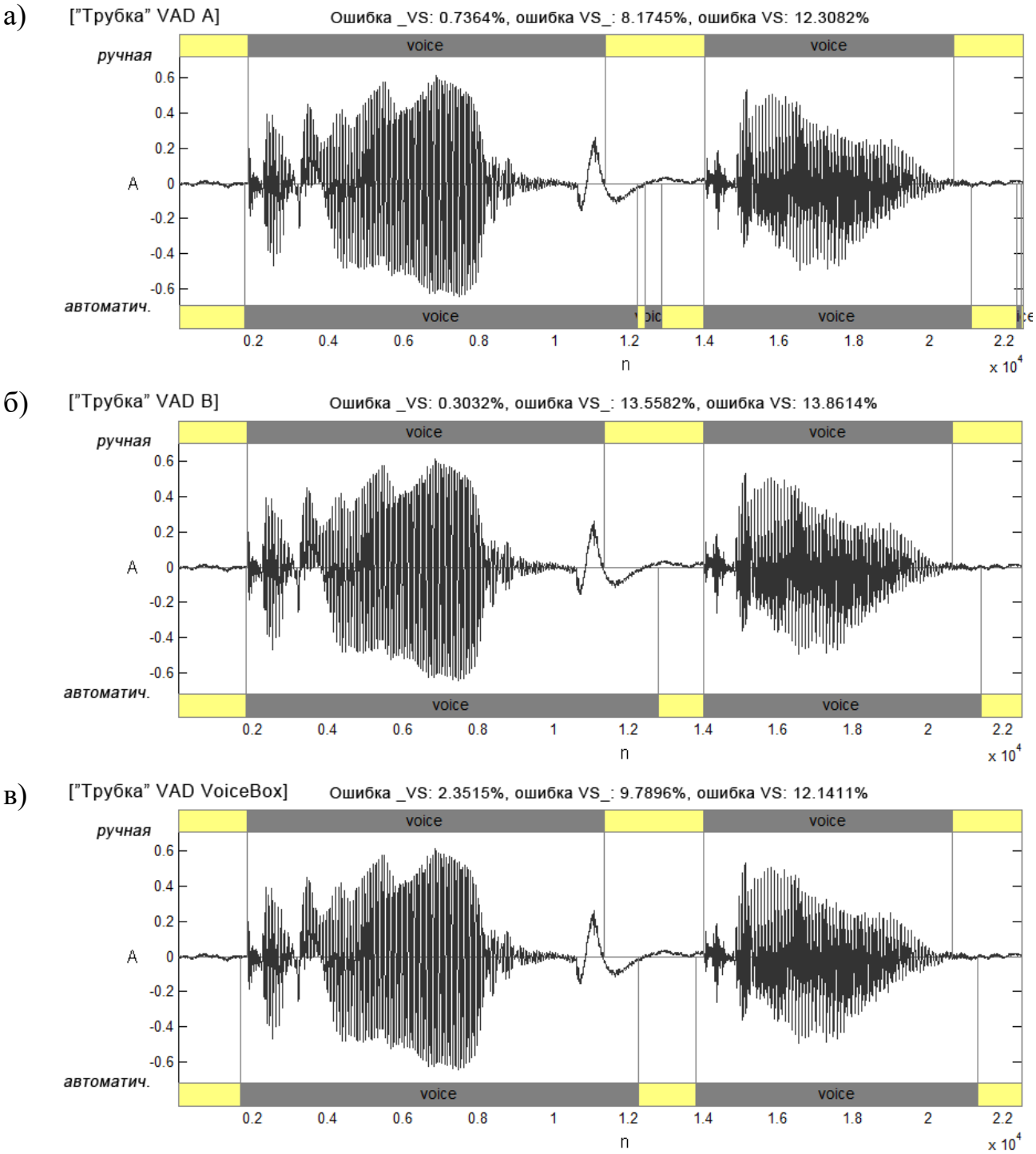
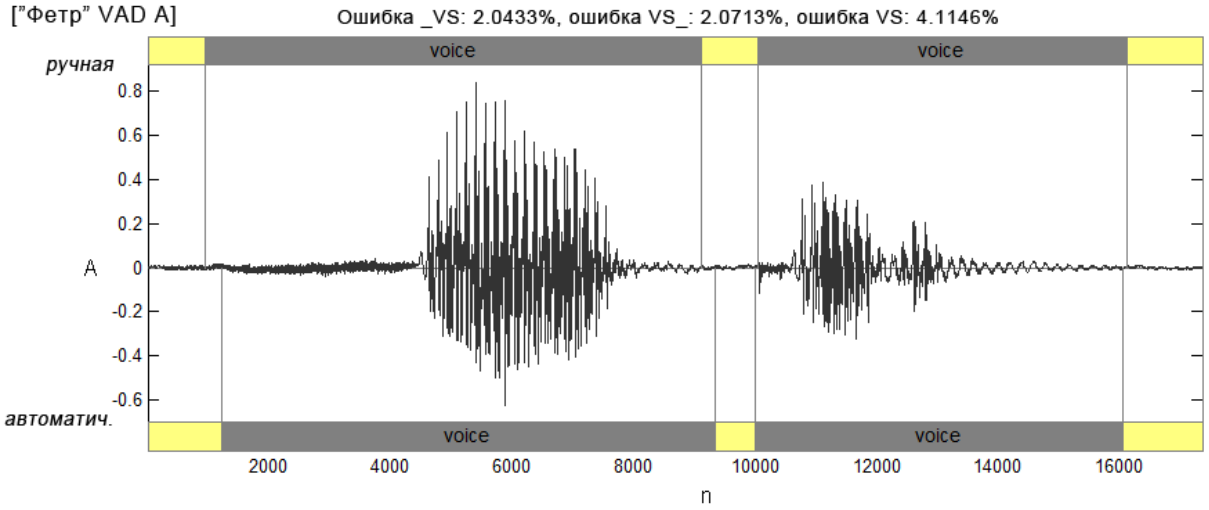
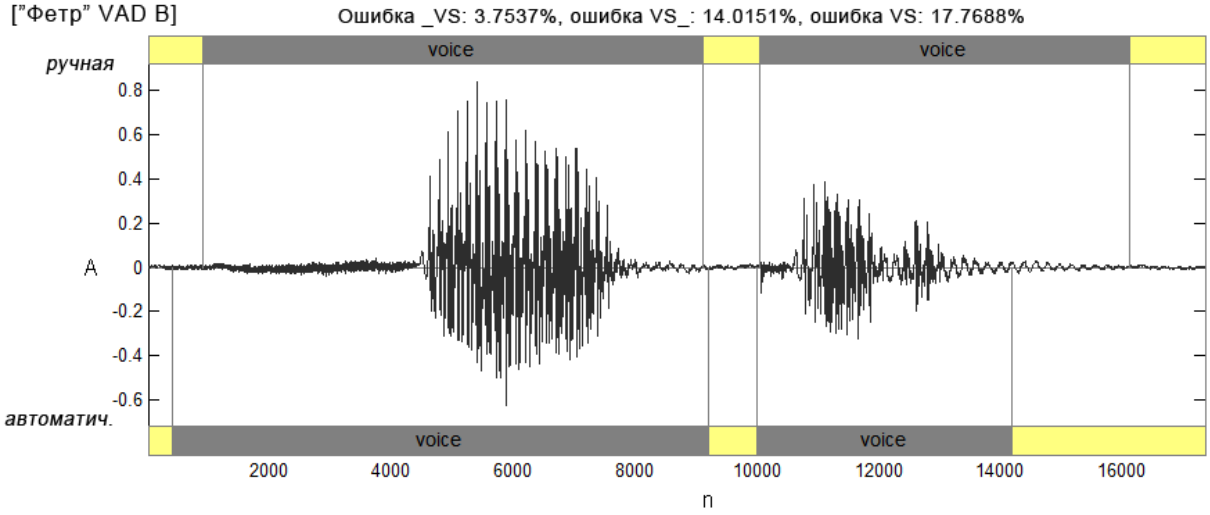


Рисунок 3.18 – Результат ручной (верхние полуоси) и автоматической (нижние полуоси) VAD-сегментации фонограммы слова «Трубка» (диктор-женщина): а) VAD-алгоритм A; б) VAD-алгоритм B; в) VAD-алгоритм VOICEBOX

а)



б)



в)

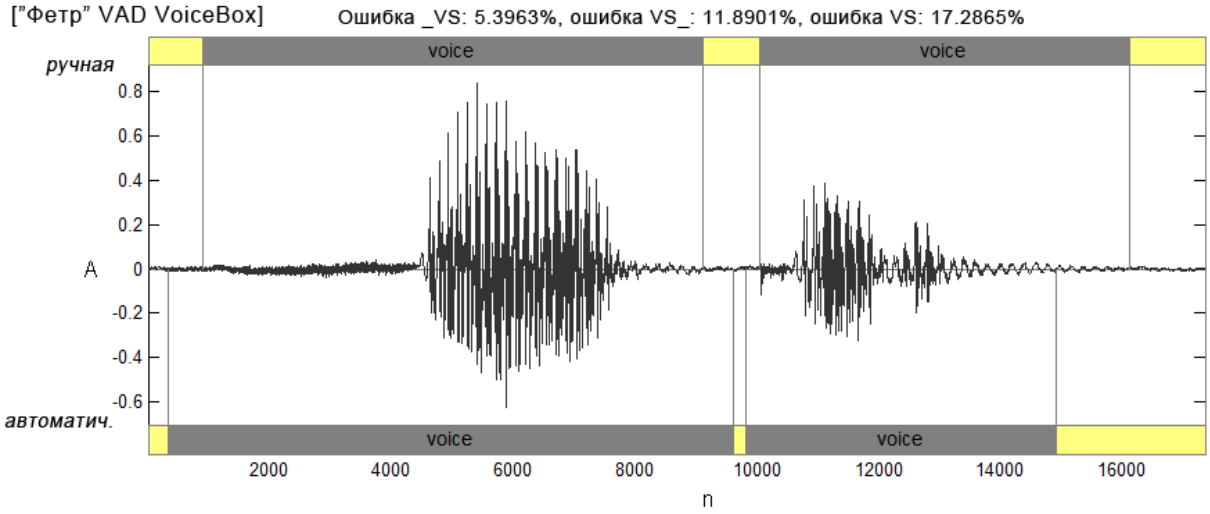


Рисунок 3.19 – Результат ручной (верхние полуоси) и автоматической (нижние полуоси) VAD-сегментации фонограммы слова «Фетр» (диктор-мужчина): а) VAD-алгоритм A; б) VAD-алгоритм B; в) VAD-алгоритм VOICEBOX

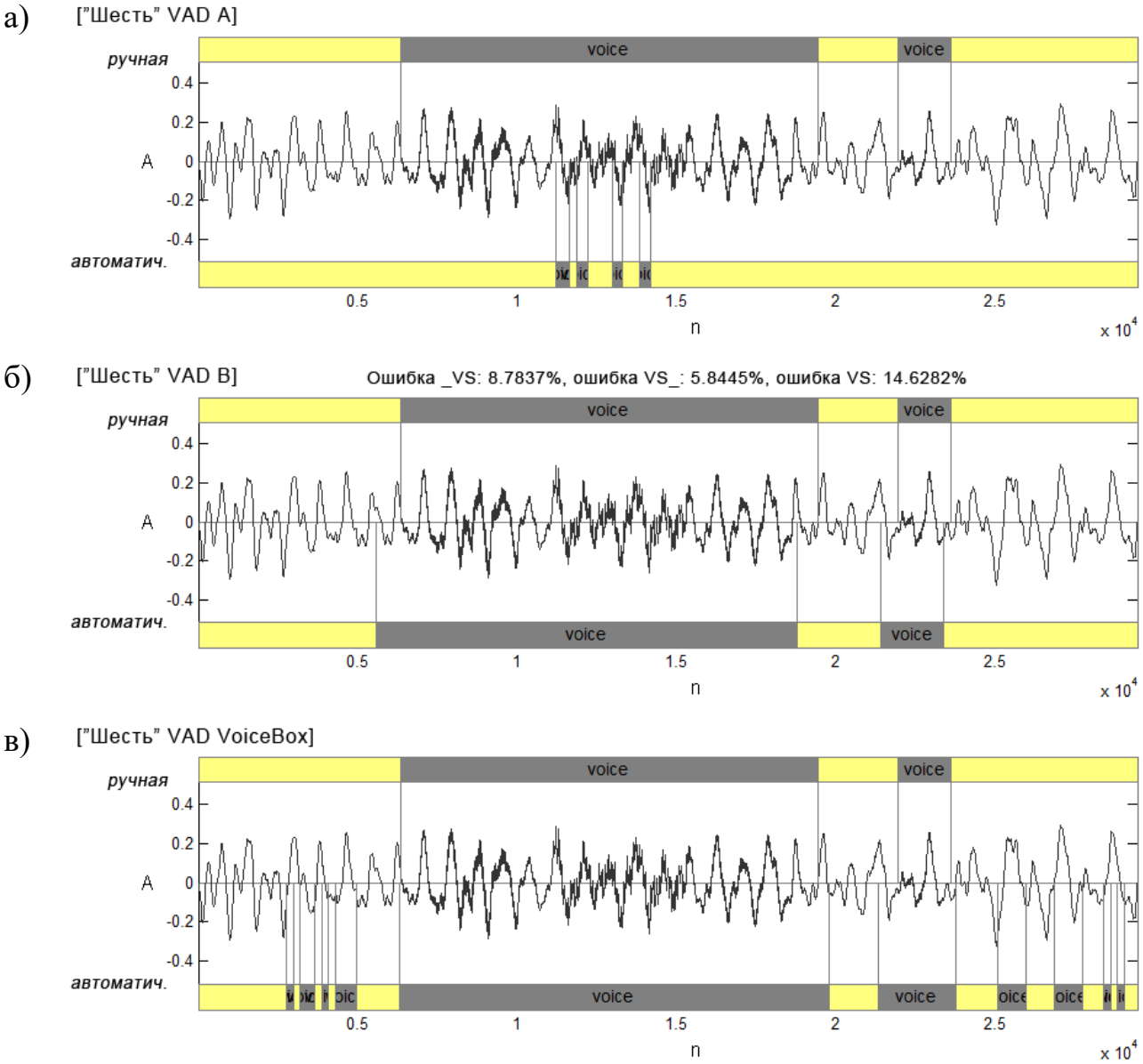


Рисунок 3.20. Результат ручной (верхние полуоси) и автоматической (нижние полуоси) VAD-сегментации фонограммы слова «Шесть» (диктор-женщина): а) VAD-алгоритм A; б) VAD-алгоритм B; в) VAD-алгоритм VOICEBOX

При рассмотрении таблиц 3.7 и 3.8 можно сделать предположение, что VAD-алгоритм B дает более точные оценки моментов начала сегментов активной речи, а также имеет низкие вероятности пропуска и ложного обнаружения речевой активности. В свою очередь, алгоритмы VAD A и VoiceBox часто дают более соответствующую субъективному восприятию при ручной сегментации оценку момента завершения речевой активности. И, наконец, алгоритм VAD VoiceBox показывает хорошую точность определения границ речевой активности

в зашумленном сигнале (как и заявлено его авторами в [127]), однако при этом дает большое число ложных сегментов активности.

Слабым местом VAD-алгоритма B являются вокализованные звуки простой квазигармонической структуры (одно переколебание на периоде ОТ, см. выше пункт 2.3.5 «Количество переколебаний на одном периоде основного тона»): как показали эксперименты, фрагменты таких звуков при сегментации могут отмечены как смычки вследствие низкой энергетической составляющей высших гармоник ОТ и сравнительно высокой частотой среза ФВЧ, превышающей частоту основной гармоники ОТ.

Для полноценного анализа эффективности сравниваемых алгоритмов исследование целесообразно повторить с использованием значительно большей размеченной базы фонограмм.

3.4.4 Ограничение остаточных колебаний вокализованных звуков перед паузой и смычкой

В случае если VAD-алгоритм не энергетический, а построен на учете, в том числе, и квазипериодического характера вокализованных звуков, правая граница звука, переходящего в паузу, может содержать сравнительно долгий фрагмент остаточных колебаний голосовых связок. Эти колебания имеют крайне низкую мощность – особенно в сравнении с предшествующей основной частью фонемы – и не воспринимаются на слух. В автоматической обработке РС такие релаксационные колебания вокализованных фонем могут приводить к искаженному вычислению ряда параметров, среди которых длительность звука, СКО отсчетов РС, средняя мощность, соотношение мощностей низкочастотной и высокочастотной составляющих и т.д.

Для корректировки правой границы звука может быть использована следующая экспериментально подтвержденная закономерность: на остаточные колебания приходится порядка *1%* энергии фонемы, несмотря на то, что по длительности она может составлять до *50*% и более от длительности фонемы.

Для определения момента разделения фонемы используется значение общей энергии E соответствующего фрагмента РС и текущего для каждого момента t интегрального значения энергии E_t:

$$E = \sum_{t=1}^{N} x_t^2,$$
$$E_t = \begin{cases} x_1^2, & t = 1 \\ E_{t-1} + x_t^2, & 2 \le t \le N \end{cases}, \qquad (3.12)$$

где x_t – отсчет сигнала в t-ый момент времени, N – количество отсчетов, представляющих реализацию фонемы.

Сегментация фонемы на активную часть и релаксационные колебания соответствует моменту превышения E_t значения kE: $E_{t-1} \le kE$ и $E_t > kE$, где k – порог, принятый в используемом алгоритме k=0,99.

На рисунке 3.21 показан пример результата работы алгоритма для звука [ъ] слова «Трубка» (диктор-мужчина). График получен при вычислении с порогом k=0,99, однако, для наглядности на рисунке уровень kE отображен ниже, на уровне k=0,97. Для предотвращения усечения алгоритмом активной части звука следует проводить проверку длительности отсекаемого фрагмента: если она составляет меньше 10% от общей длительности звука, операция отменяется.

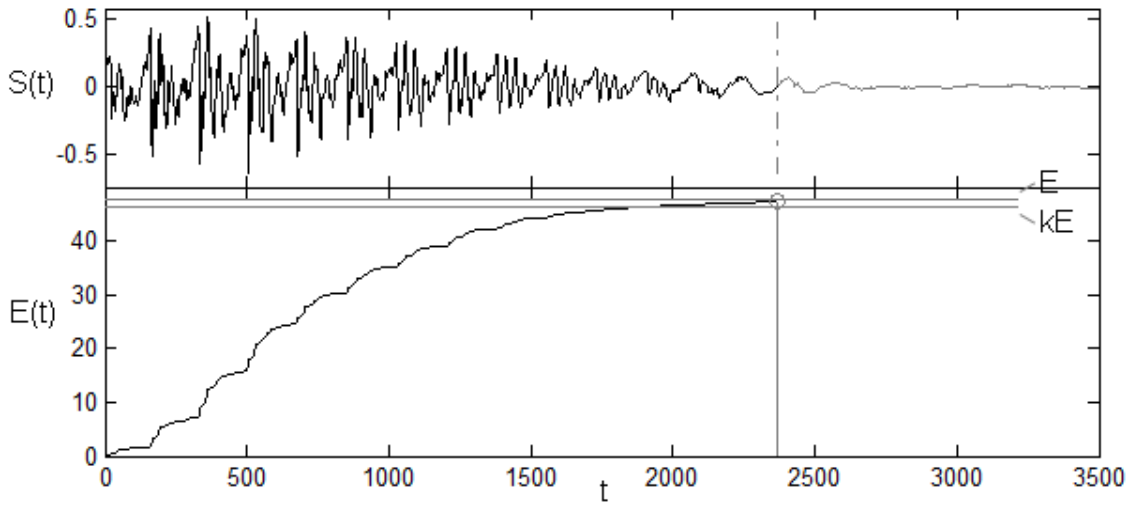


Рисунок 3.21 – Ограничение остаточных колебаний в звуке перед паузой: осциллограмма *S(t)* целиком звука; накопление текущей энергии *E(t)* и порог

3.5 Сегментация второго уровня: выделение типовых фрагментов речи

Задача выделения типовых фрагментов речи состоит в определении временных границ вокализованных, взрывных и шумных звуков на интервалах речевой активности в РС.

3.5.1 Принципы обработки

Разделение активных участков речи на шумные, вокализованные и взрывные производится на основе выявления характерных для определенных типов признаков. К примеру, такую сегментацию можно осуществить путем формирования выборочных корреляционных функций фильтрованной от низкочастотных наводок сигнальной последовательности на малых окнах оценивания. Для сегментов разных типов такие корреляционные функции различаются, можно выделить три характерных поведения КФ: $K_{ш}(\tau)$ – корреляционная функция для сегмента шумных звуков; $K_{вз}(\tau)$ – корреляционная функция взрывного сегмента; $K_{вок}(\tau)$ – корреляционная функция вокализованного сегмента. Структура обработки фонограммы при таком подходе имеет вид, показанный на рисунке 3.22.

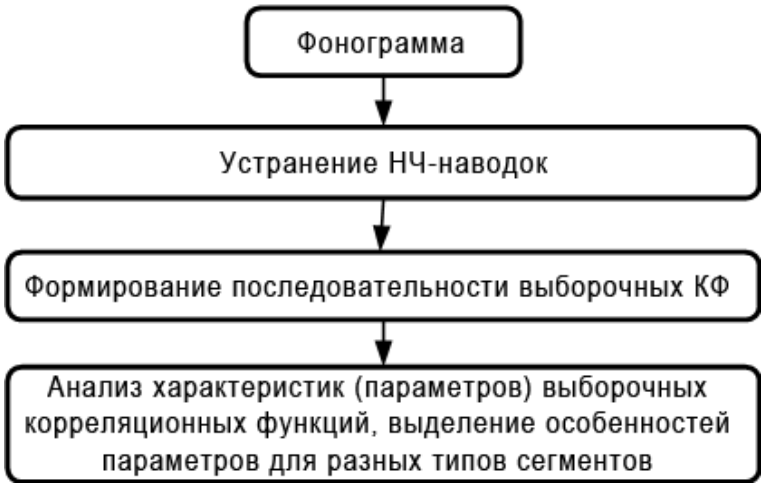


Рисунок 3.22. Последовательность действий при обработке фонограммы для разбиения на сегменты «шумный/вокализованный/взрывной»

На практике наиболее удобным является разделение активных участков речи на три перечисленных основных типа в два этапа. Первым этапом может выступать разделение «шумный/нешумный», тогда вторым этапом будет выступать сегментация нешумных фрагментов на вокализованные и взрывные. Соответственно, в ином подходе на первом этапе можно производить сегментацию «вокализованный/невокализованный», тогда на втором сегментировать невокализованные участки на фрагменты «шумный/взрывной».

3.5.2 Алгоритм сегментации «шумный/нешумный»

В разработанном алгоритме выделения из сегментов речевой активности шумных звуков пороговым параметром принятия решения является отношение стандартного отклонения от линии нуля отсчетов РС, пропущенного через ФНЧ, призванный сгладить выбросы сигнала в шумных фрагментах, к стандартному отклонению от линии нуля отсчетов исходного речевого сигнала. Для выполнения этой операции сначала речевой сигнал пропускается через ФНЧ с апертурой прямоугольного окна около *0,8* мс. Затем по активным участкам речи перемещается окно оценивания длиной N отсчетов (порядка *10* мс), в пределах которого осуществляется оценка стандартного отклонения для фильтрованного сигнала y_F и исходного сигнала y:

$$\begin{aligned} \sigma_1 &= \sqrt{\frac{1}{N-1}\sum_{n=1}^{N} y_F^2(n)} \\ \sigma_2 &= \sqrt{\frac{1}{N-1}\sum_{n=1}^{N} y^2(n)} \end{aligned} \qquad (3.13)$$

Далее вычисляется отношение σ_1/σ_2, по которому и принимается решение: относится ли фрагмент речи, находящийся в текущем окне оценивания, к шумному или нет. Для шумных участков отношение σ_1/σ_2 заведомо меньше единицы (рисунок 3.23, $\sigma_1/\sigma_2 = 0.23$). В то же время на вокализованных и взрывных участках речи фильтрованный сигнал будет успевать «следить» за

траекторией исходного, поэтому отношение двух стандартных отклонений в таких случаях будет несильно отличаться от единицы в большую или меньшую сторону (рисунок 3.24, $\sigma_1/\sigma_2 = 0.91$).

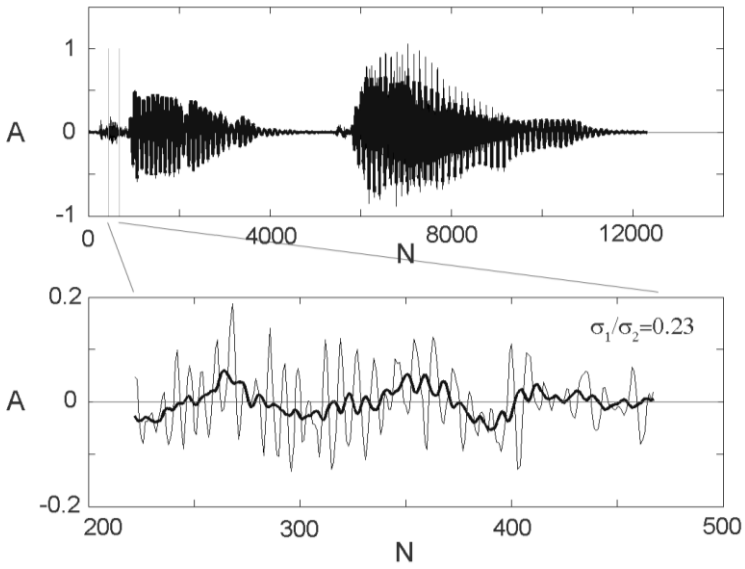


Рисунок 3.23 – Пример определения шумного фрагмента речи

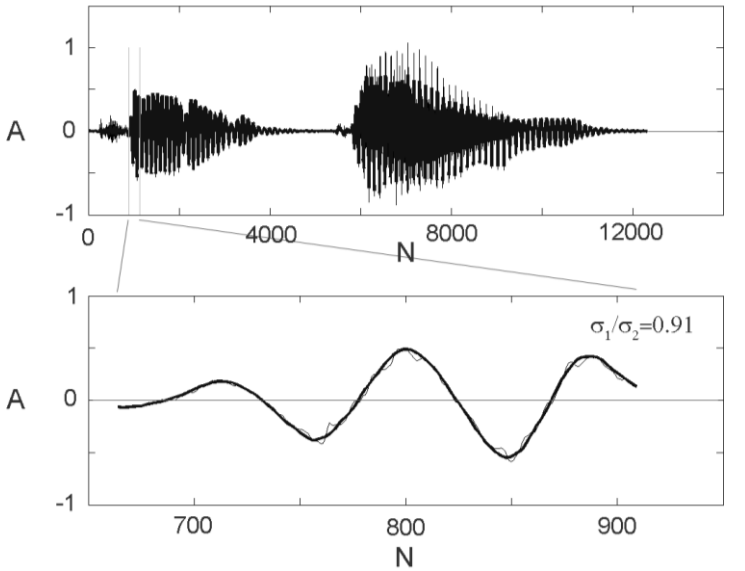


Рисунок 3.24 – Пример определения «нешумного» фрагмента речи

3.5.3 Алгоритм сегментации «вокализованный/невокализованный»

Вокализованные звуки характерны активным участием в их образовании голосовых связок диктора, совершающих при этом колебания с частотой ОТ. В зависимости от состояния речевого аппарата и его эволюций в процессе речи, в

РС меняется в некоторых пределах как частота ОТ, так и структура сигнала в пределах одного периода ОТ. Таким образом, квазипериодическая структура сигнала вокализованных звуков является значимым критерием для разделения сигнала на сегменты «вокализованный/невокализованный».

Кроме того, как показывают результаты исследования фонограмм, вокализованные звуки отличает сравнительно большая энергия и длительность (однако, большая длительность может также наблюдаться у шипящих согласных).

Наличие во фрагменте РС вокализации может быть определено во временной области с помощью анализа локальных экстремумов. В корреляционной обработке вокализованные фонемы, как квазипериодические сигналы, будут давать квазипериодическую затухающую корреляционную функцию [128]. Таким образом, по характеру поведения КФ также может быть принято решение о вокализации некоторого фрагмента РС. В спектральной области вокализованные фрагменты будут содержать пики на частотах, кратных частоте ОТ. Для увеличения вероятности правильной сегментации «вокализованный/невокализованный» при применении спектрального анализа может использоваться накопление некоторой целевой функцией, которая для вокализованного звука указанная целевая функция будет иметь ярко выраженный пик на частоте ОТ.

Как было сказано выше в пункте 2.3.2, посвященном исследованию энергетических информативных параметров звуков, наиболее прямым и простым параметром, отражающим наличие или отсутствие вокализации, является средняя мощность фрагмента. Так, энергия вокализованных звуков на спектрограмме сосредотачивается в области низких частот (всилу относительно низкой частоты ОТ). Поэтому для улучшения результатов работы алгоритма сегментации «вокализованный / невокализованный» на основе параметра средней мощности необходимо выделять его только для низкочастотной составляющей сигнала. Учитывая, что верхняя граница диапазона частот ОТ для женского голоса составляет *500* Гц (см. выше пункт 1.1.3 «Произнесение и восприятие речи

человеком. Фонетическое строение сигнала русской речи»), приемлемой частотой среза НЧ-фильтра можно считать *750* Гц.

В целом, задачи сегментации «вокализованный/невокализованный» и сегментации вокализованных фрагментов на периоды ОТ имеют общие подходы к решению. Зачастую данные два вида сегментации осуществляются в рамках одного этапа обработки, в котором алгоритм определения наличия вокализации в процессе своей работы получает оценку частоты ОТ.

Ниже приводится описание разработанной реализации корреляционного алгоритма, выполняющего сегментацию «вокализованный / невокализованный».

В реализации алгоритма вычисляются взаимные корреляционные функции (ВКФ) фрагмента фонограммы, находящегося в текущем окне оценивания, и фрагментов такой же длины, находящихся во временной области до и после исследуемого. Такой подход позволяет более точно найти позиции начала и окончания вокализованного сегмента, нежели при вычислении АКФ, требующей более широкого окна оценивания. Необходимо отметить, что при описанном вычислении ВКФ в общем случае ее максимум не будет приходиться на нулевой сдвиг, так как два сравниваемых вокализованных участка в общем же случае будут иметь различные начальные фазы.

В качестве признака вокализации текущего фрагмента используется квазипериодический затухающий характер поучаемой корреляционной функции (рисунок 3.25). Алгоритмом оценивается периодичность следования экстремумов ВКФ: определяются локальные максимумы и минимумы полученной ВКФ, оценивается математическое ожидание *M* интервала их следования, оценивается стандартное отклонение *SD* интервалов между экстремумами от этого математического ожидания. Для оценки относительной степени периодичности следования экстремумов используется отношение оценок стандартного отклонения и математического ожидания.

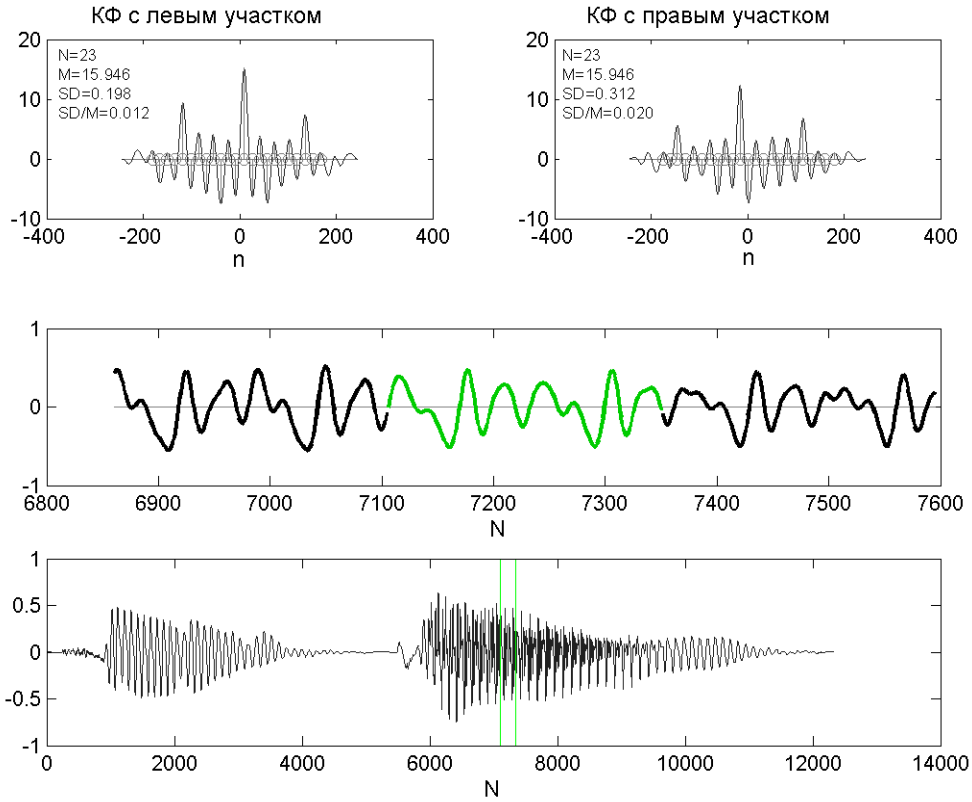


Рисунок 3.25. ВКФ смежных вокализованных фрагментов РС: основной анализируемый фрагмент подсвечен зеленым

При оценке перечисленных параметров следует рассматривать не весь интервал ВКФ, а только приблизительно ¾ ее значений, расположенных в центральной части, так как по краям, где области перекрытия сравниваемых отрезков уже небольшие, велика вероятность возникновения паразитных максимумов и минимумов. Математическое ожидание следует определять не непосредственно по интервалу следования абсцисс максимумов и минимумов, а путем деления общей длительности исследуемого интервала ВКФ к числу найденных экстремумов, приходящихся на данный интервал. Этот метод позволяет отсеять случаи, когда экстремумы находятся на достаточно регулярных расстояниях друг от друга, но распределены неравномерно по длине ВКФ (рисунок 3.26).

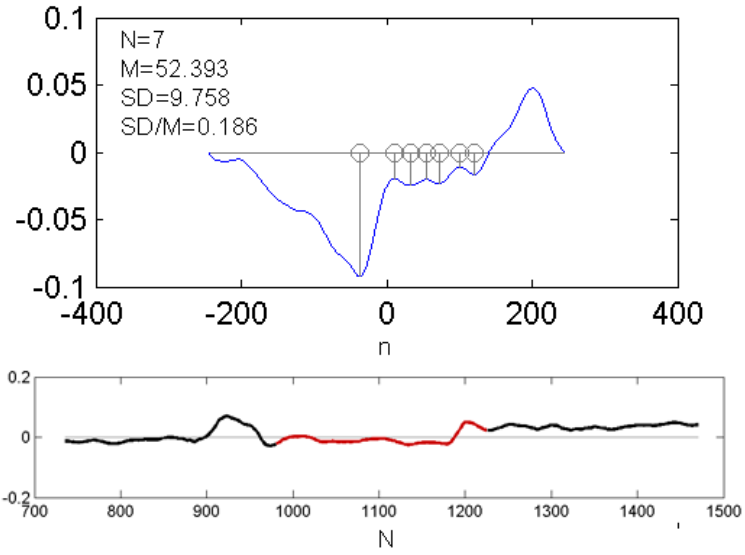


Рисунок 3.26 – ВКФ (верхний график) и фрагменты РС, по которым она получена (красный и черный справа от красного, нижний график)

3.6 Третий уровень сегментации: сегментация на периоды основного тона

Задача сегментации на периоды основного тона заключается в определении временных границ отдельных колебаний РС, соответствующих колебаниям голосовых связок диктора, на интервалах вокализованных звуков.

Как уже отличалось ранее, неотъемлемой особенностью вокализованных звуков является их квазипериодическая структура, что позволяет решать данную задачу сегментации непосредственно во временной области.

3.6.1 Реализация корреляционного алгоритма ОТ-сегментации

При работе корреляционного алгоритма из речевого фрагмента с помощью прямоугольного временного окна выделяется участок длительностью не менее двух максимально возможных для человеческого голоса периодов ОТ (для человека T_{MAX} имеет значение порядка 20 мс) симметрично относительно известной (или предполагаемой – на первом шаге работы алгоритма) левой границы периода ОТ.

В АКФ этого фрагмента находится максимум, лежащий в пределах возможного периода ОТ человека, то есть между значениями 2,5 мс и 20 мс

(рисунок 3.27б). Смещение, соответствующее этому максимуму, и является оценкой среднего для текущего фрагмента периода ОТ.

Так как получаемая оценка является величиной усредненной, она нуждается в коррекции: правая граница периода ОТ вычисляется прибавлением к левой границе длительности оценки среднего периода ОТ с коррекцией до точки ближайшего пересечения нуля в направлении снизу вверх (для сегментации предлагается руководствоваться переходами сигнала через ноль в направлении снизу вверх) пропущенного через ФНЧ речевого сигнала. Завершающим шагом коррекции должен стать поиск ближайшего к положению найденной границы перехода через ноль снизу вверх уже исходным сигналом. На рисунке 3.27а левая серая вертикальная черта – известная граница начала периода ОТ, правая серая черта – полученная непосредственно по оценке среднего периода ОТ правая граница периода ОТ. Черная черта – скорректированная правая граница.

Таким образом, на рисунке 3.27а интервал между левой серой чертой и черной чертой является оценкой текущего периода ОТ, а интервал между серыми чертами – средним периодом ОТ исследуемого фрагмента РС.

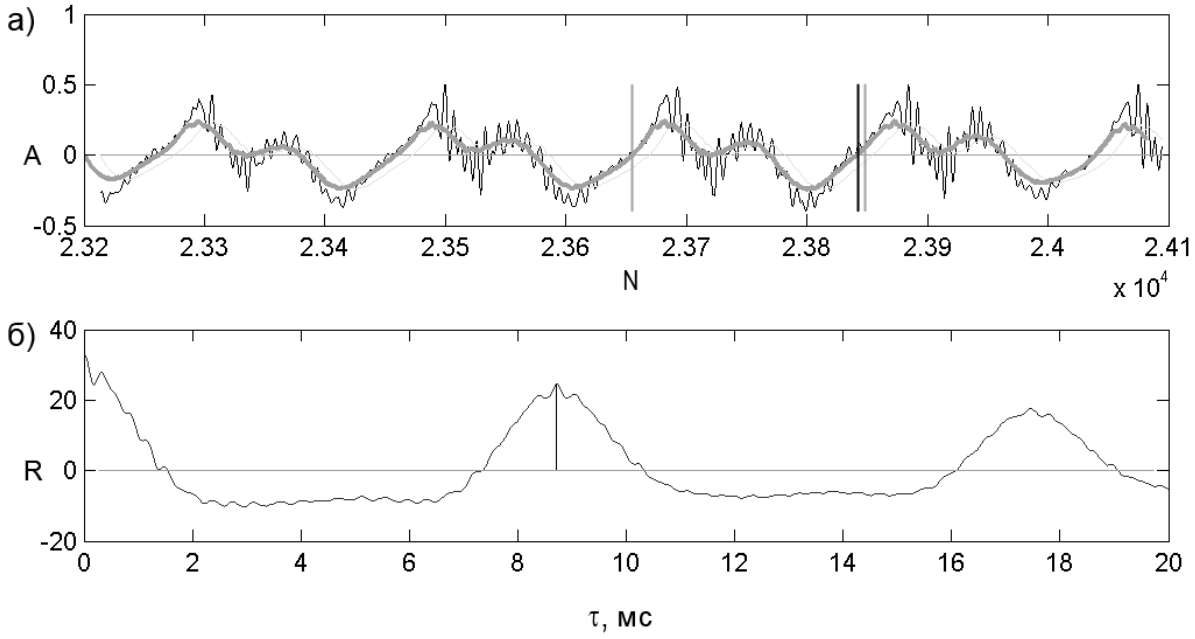


Рисунок 3.27. Выделение периодов ОТ: а) исследуемый фрагмент фонограммы; б) автокорреляционная функция исследуемого фрагмента

В завершение стоит отметить, что алгоритм, построенный на определении периода ОТ по АКФ участка фонограммы может быть использован для сегментации на субпериоды шумных фрагментов и фрагментов паузы: в таких участках фонограмма будет разбита на отрезки, края которых соответствуют точкам перехода сигнала через уровень нуля (в заранее определенном направлении: снизу вверх или сверху вниз). Такая сегментация, в частности, может применяться в алгоритмах модификации параметров произнесения РС (см. ниже подраздел 4.7 «Модификация произнесения речи»).

3.6.2 Разработка алгоритма ОТ-сегментации во временной области

Предлагаемый алгоритм сегментации вокализованных фрагментов основан на подходе, описанном выше в пункте 3.2.1 «Алгоритм выделения огибающей»: в обоих случаях алгоритмами производится отбор локальных максимумов во временном представлении РС.

В алгоритме ОТ-сегментации (рисунок 3.28) моменты времени отбираемых локальных максимумов должны соответствовать границам ОТ-сегментов сигнала. Для этого в алгоритм помимо параметра L, ограничивающего в данном случае максимально возможный период ОТ человеческого голоса, вводится также параметр, ограничивающий минимально возможный период ОТ – данные два параметра обозначены соответственно как L_{max} и L_{min}. Кроме того, локальные максимумы, являющиеся кандидатами на границу ОТ-сегмента, проходят ряд дополнительных условий, направленных на борьбу с пропусками границ ОТ-сегментов и образующих в алгоритме внутренний цикл (см. рисунок).

Одной из трудностей, с которой может сталкиваться данный алгоритм – влияние внутренних переколебаний в пределах периода ОТ, которые могут иметь достаточный размах, чтобы порождать ложные отметки границ сегментов. В связи с этим перед подачей сигнала на вход алгоритма необходимо провести соответствующую корректировку полярности РС. К слову, данная корректировка требуется также и для упоминаемого далее стороннего алгоритма SEDREAMS.

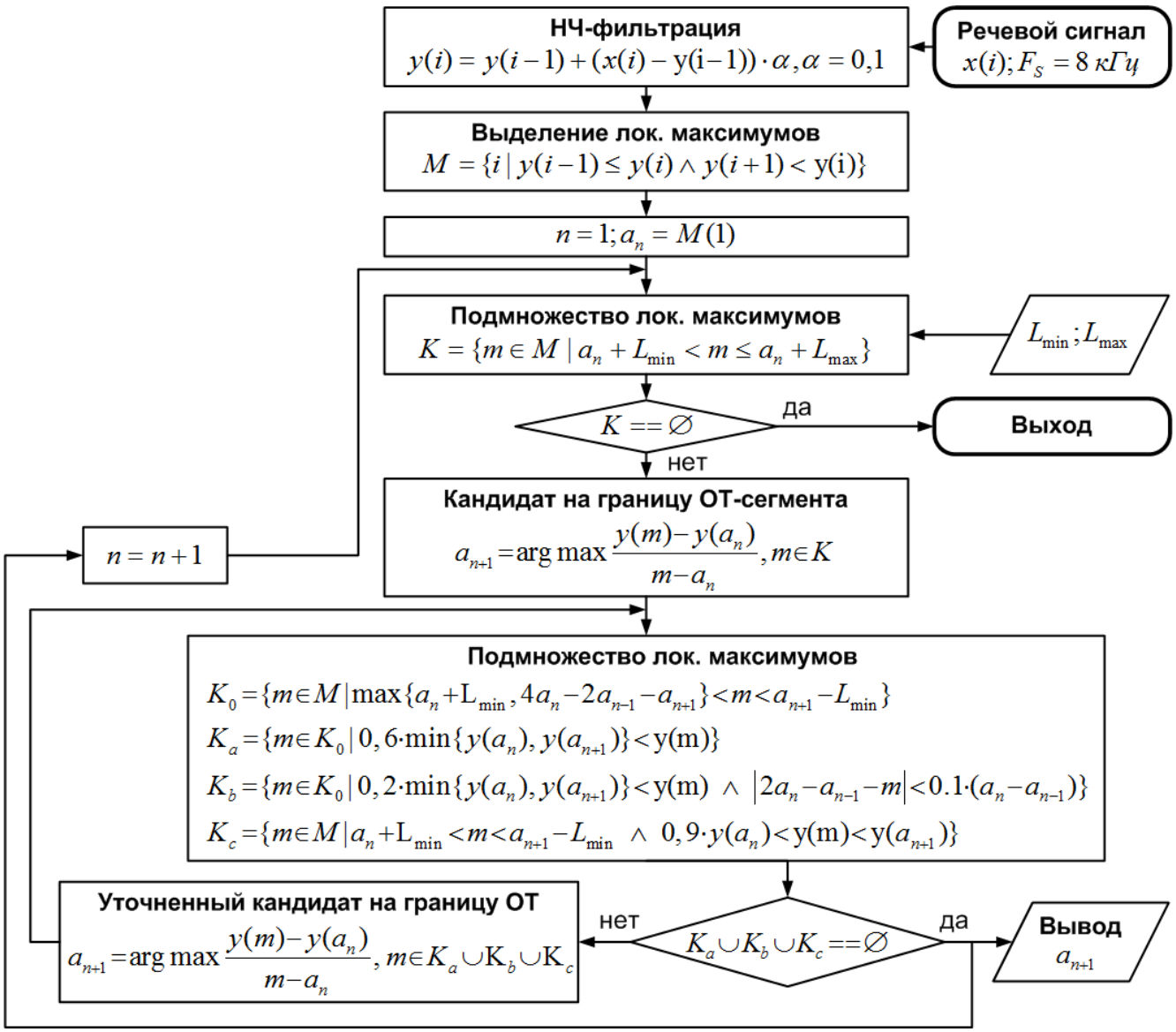


Рисунок 3.28 –Алгоритм ОТ-сегментации во временной области

Эффективность разработанного алгоритма (далее обозначен как алгоритм EOT) проанализирована в сравнении с рядом других известных современных алгоритмов ОТ-сегментации: DYPSA, SEDREAMS и SE-VQ [129-131]. Экспериментальное исследование производилось на фонограммах базы TIDIGITS с использованием в качестве вспомогательных алгоритмов оценки частоты ОТ PEFAC [122] и RAPT [84], а также алгоритма SRH [132] оценки степени вокализованности фрагментов РС. Исследование производилось при воздействии на РС естественных помех различной природы: шум в салоне автомобиля, шум в салоне поезда, шум улицы вблизи автодороги, шум толпы.

С помощью вспомогательных алгоритмов оценки частоты ОТ PEFAC и RAPT на интервале анализа определяется длительность текущего периода колебания голосовых связок. Для дальнейшей работы отбираются только интервалы, в которых оценки частоты ОТ, получаемые от данных двух алгоритмов, оказываются в достаточной степени равными (допустимая разница оценок около 0,5 мс). Алгоритм SRH оценки наличия вокализации используется для исключения из анализа невокализованных фрагментов. Наконец, в качестве эталонной длительности периода ОТ применяются оценки, полученные от PEFAC и RAPT (конкретная из двух оценок выбирается в пользу анализируемых алгоритмов). Алгоритмы PEFAC, RAPT и SRH всегда работают с чистыми фонограммами, а анализируемые алгоритмы ОТ-сегментации – с фонограммами с соответствующими экспериментам ОСШ.

На рисунках 3.29-3.31 приведены графики полученных оценок показателей надежности алгоритмов ОТ сегментации:

– *IDR* (identification rate) – частота правильного определения сегментов: доля периодов ОТ, для которых определена ровно одна граница ОТ;
– *MR* (miss rate) – частота пропуска сегментов: доля периодов ОТ, для которых не найдено ни одной границы ОТ, ошибка первого рода;
– *FAR* (false alarm rate) – частота сегментов с ложными границами (более одной границы ОТ), ошибка второго рода.

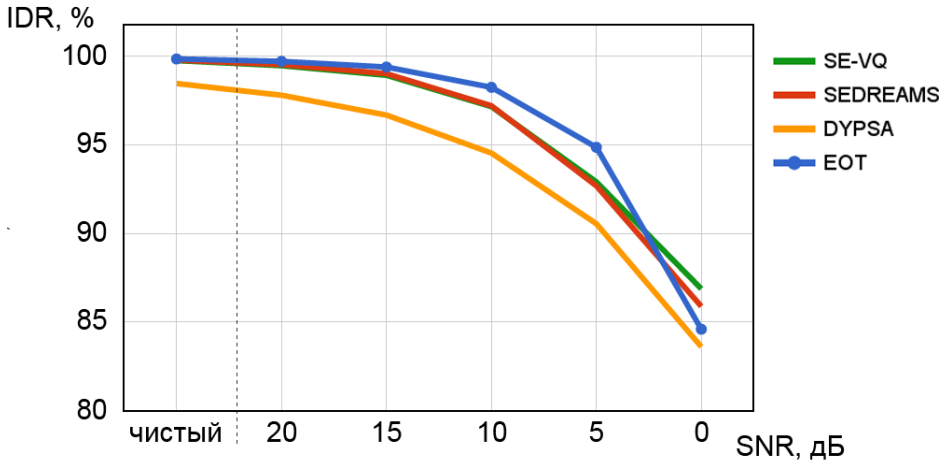


Рисунок 3.29 – Частота правильного определения ОТ-сегментов

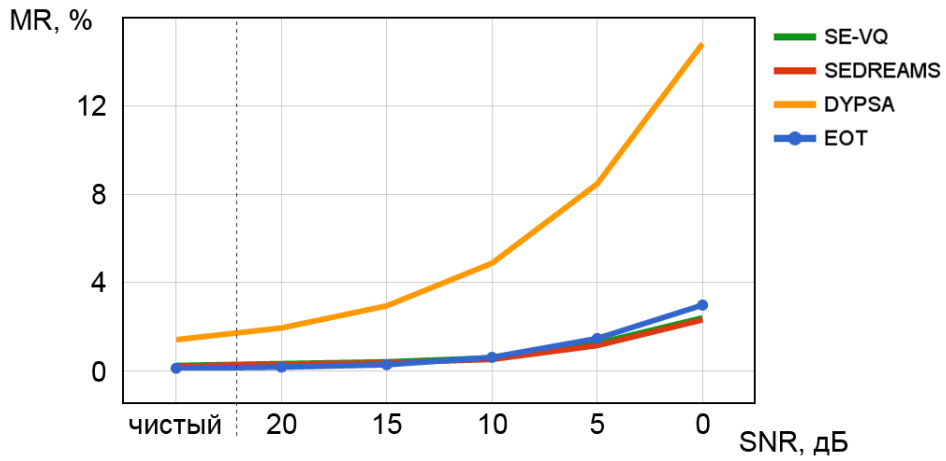


Рисунок 3.30 – Частота пропуска ОТ-сегментов

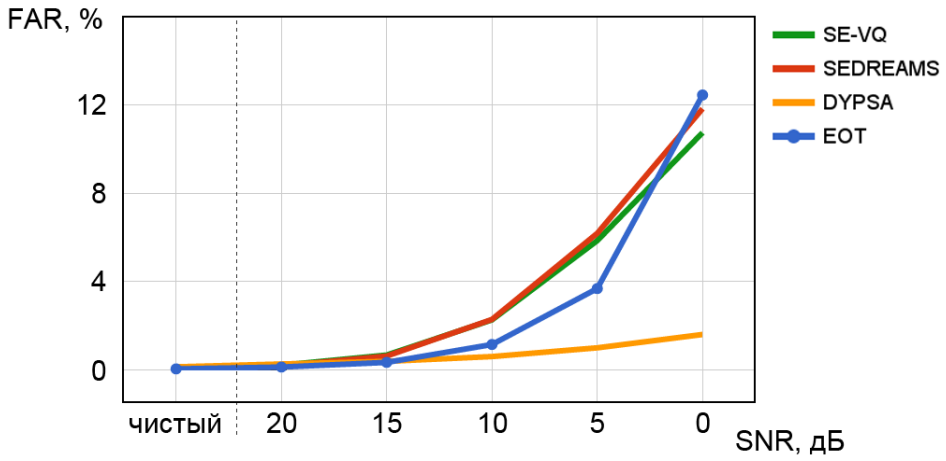


Рисунок 3.31 – Частота ОТ-сегментов, содержащих ложные отметки границ

По показателям надежности разработанный алгоритм EOT в экспериментах показал значения лучше, чем у всех аналогов, при ОСШ не менее 15 дБ. В целом же, значения по алгоритму EOT лежат в рамках аналогичных значений сторонних алгоритмов до ОСШ 5 дБ.

На рисунке 3.32 показаны значения, полученные для показателя точности ОТ-сегментации: *IDA* (identification accuracy) – стандартное отклонение ошибки определения временной границы периода ОТ. Разработанный алгоритм не выходит за рамки значений *IDA* сторонних алгоритмов при ОСШ не менее 5 дБ.

Наконец, на рисунке 3.33 приведены результаты оценки показателя скорости работы сравниваемых алгоритмов *SF* (speed factor), определяемого как отношение времени обработки всех анализируемых фонограмм *TPT* (total processing time) к их общей длительности *SSD* (source signal duration):

$$SF = \frac{TPT}{SSD} \times 100\%. \qquad (3.14)$$

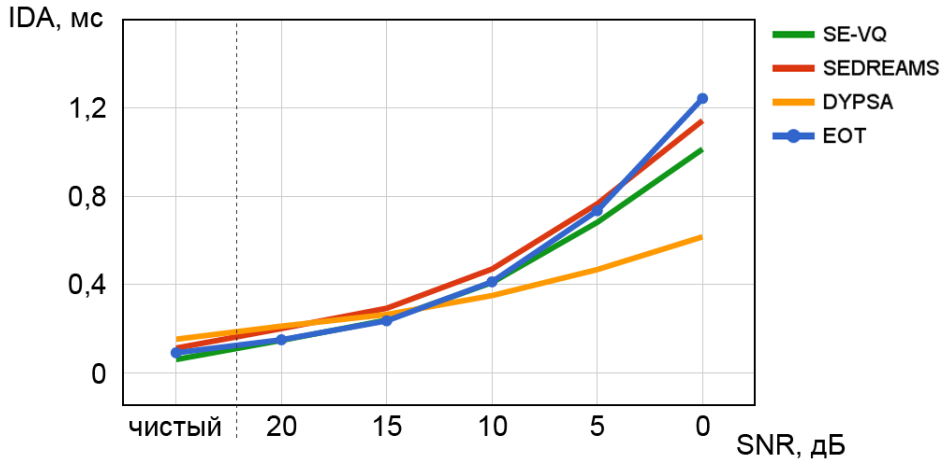


Рисунок 3.32 – Стандартное отклонение ошибки определения временной границы периода ОТ

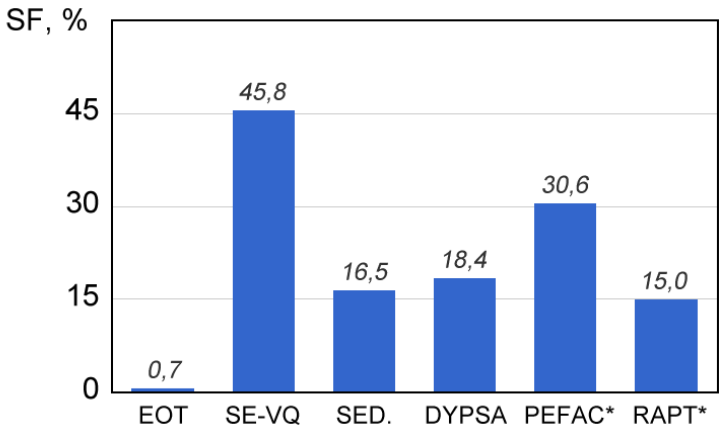


Рисунок 3.33 – Показатели скорости алгоритмов ОТ-сегментации

Важным достоинством предложенного алгоритма является обработка РС только во временной области без применения математических преобразований высокой вычислительной сложности. В результате алгоритм обладает крайне высоким быстродействием, в десятки раз опережая аналоги.

Таким образом, разработанный алгоритм ОТ-сегментации EOT позволяет эффективно заменить рассмотренные известные аналоги при обработке РС с ОСШ не менее 5 дБ, потребляя при этом значительно меньше вычислительных ресурсов.

3.6.3 Анализ трендов и разладок для определения границ вокализованных звуков

Как было сказано выше в подразделе 3.2.2 «Применение огибающей в выявлении переходных участков фонограммы», сегментация по огибающей имеет высокую вероятность пропуска границы между вокализованными звуками с плавным по амплитуде переходом. Для выявления таких переходов необходимо учитывать изменения в структуре периодов ОТ РС.

Для характеризации изменений структуры периодов ОТ возможно ввести метрики, учитывающие разницу энергетик и низкочастотных структур двух сравниваемых периодов ОТ. При этом плавные переходы внутри одного звука и между звуками будем называть трендами, а сравнительно резкие переходы между звуками – разладками. Можно предложить два вида метрик – амплитудно-структурную и структурную.

Амплитудно-структурные тренды выражаются как в амплитудных, так и в структурных изменениях периодов ОТ:

$$M_{амп-стр}(j) = \frac{1}{T_{OT}(j) A_{ср}(j)} \sum_{i=1}^{T_{OT}(j)} \left| x(i, j) - x(i, j-1) \right|, \qquad (3.15)$$

где $M(j)$ – последовательность метрик, полученная на последовательности периодов ОТ $T_{от}(j\text{-}2) \Rightarrow T_{от}(j\text{-}1) \Rightarrow T_{от}(j) \Rightarrow T_{от}(j+1) \Rightarrow T_{от}(j+2) \Rightarrow \ldots$; $T_{от}(j)$ – длительность j-го периода ОТ; $A_{ср}(j)$ – средняя амплитуда группы периодов ОТ (группа состоит из 2...3 периодов), вычисленная на скользящем временном интервале для j-го периода ОТ.

Структурные тренды выражаются в структурных изменениях периодов ОТ:

$$M_{стр}(j) = \frac{1}{T_{OT}(j)} \sum_{i=1}^{T_{OT}(j)} \left| \frac{x(i, j)}{x_{\max}(j)} - \frac{x(i, j-1)}{x_{\max}(j-1)} \right|, \qquad (3.16)$$

где $x_{max}(j)$ – максимальный отсчет на интервале j-го периода ОТ.

По приведенным выше формулам (3.15) и (3.16) видно, что для вычисления предложенных метрик размерности периодов ОТ (длины представляющих их векторов отсчетов РС) должны быть равны. Однако в общем случае в реальном РС длины двух смежных векторов ОТ-кластеров отличаются на несколько отсчетов. Поэтому перед вычислением метрик необходимо уравнять длины двух исследуемых кластеров путем векторной интерполяции.

Функциональная схема анализа РС с целью обнаружения разладок и трендов временной функции показана на рисунке 3.34:

- анализ и разбиение вокализованных сегментов на периоды ОТ (блоки 1 и 2, рисунок 3.34);
- векторная интерполяция (блоки 5 и 9) наименьшего из двух анализируемых кластеров ОТ с целью приведения векторов отсчетов к одинаковой размерности, когда $T_{om}(j) \neq T_{om}(j\text{-}1)$;
- вычисление значений метрик по соседним ОТ-кластерам для анализируемого участка фонограммы (блоки 6 и 10);
- анализ последовательности значений метрик (блок 11);
- ФНЧ и ФВЧ (с частотами среза порядка *300...500* Гц, блоки 3 и 4, 7 и 8) позволяют анализировать более детально изменения в структуре низкочастотной, формантной и шумной компонентах, что делает возможным выполнение сегментации при смене одного типового сегмента другим.

Векторная интерполяция периодов ОТ с целью уравнивания их длительности может быть произведена по формулам.:

$$\begin{aligned} &\tilde{y}(i) = \left(y\left(\left[\tilde{x}(i)\right]+1\right) - y\left(\left[\tilde{x}(i)\right]\right)\right)\left(\tilde{x}(i) - \left[\tilde{x}(i)\right]\right) + \mathrm{y}\left(\left[\tilde{x}(i)\right]\right), \\ &\tilde{x}(i) = 1 + (i-1)\frac{N_1 - 1}{N_2 - 1}, \quad i = \overline{1, N_2}, \end{aligned} \tag{3.17}$$

где *y(i)* – значение *i*-го отсчета исходного вектора периода ОТ; *ỹ(i)* – значение i-го отсчета интерполированного вектора; N_1, N_2 – соответственно исходная и целевая длина вектора отсчетов периода ОТ; квадратными скобками [ ] показана операция взятия целой части дробного числа.

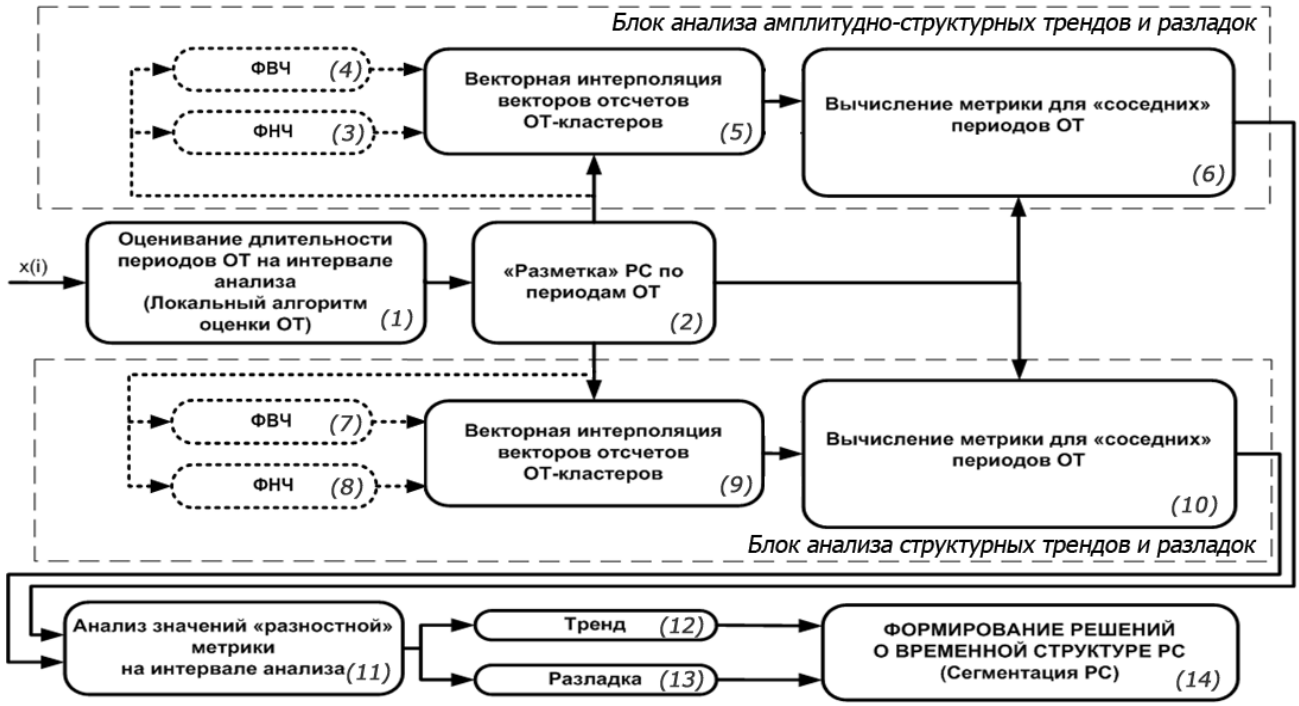


Рисунок 3.34 – Обработка речевого сигнала с целью обнаружения разладок и трендов временной функции

После интерполяции длительности обоих векторов отчетов равны N_2. Стоит отметить, что $x_1(i)$ – целые числа, а $x_2(i)$ – в общем случае дробное.

На рисунке 3.35 показан пример амплитудно-структурной метрики, вычисленной по формуле (3.15), на примере фонограммы слова «Сидорова», которое начинается с шумного звука [с], а оставшаяся часть слова состоит только из вокализованных звуков. Пример результата вычисления по формуле (3.16) структурной метрики показан на рисунке 3.36 для этого же речевого сигнала. На рисунках 3.35 и 3.36 метрики отображены в виде вертикальных линий с маркером на конце. Высота линии пропорциональна величине метрики. Метрики ставятся в точках, разделяющих исследуемые смежные периоды ОТ.

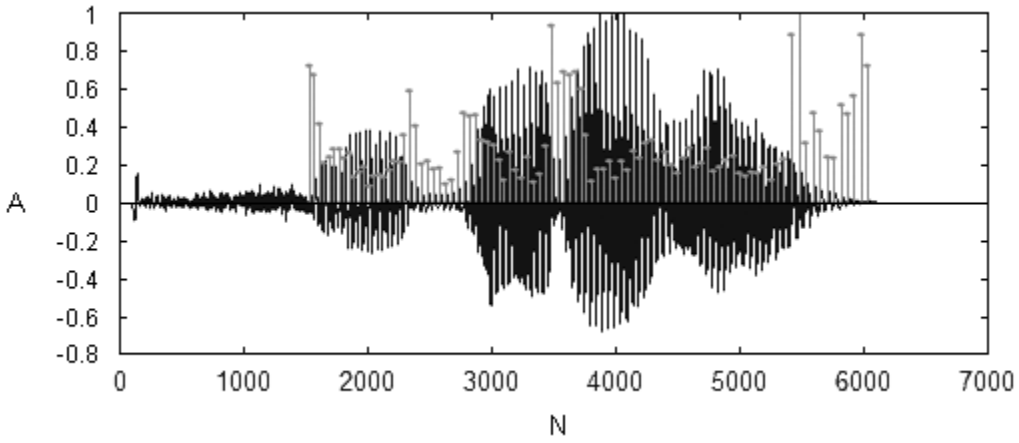


Рисунок 3.35 – Пример амплитудно-структурной метрики

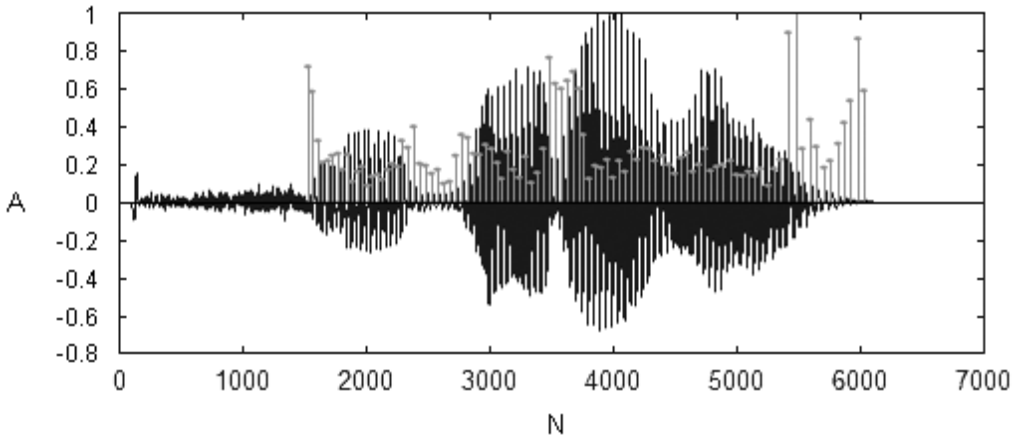


Рисунок 3.36 – Пример структурной метрики

При сравнении рисунков 3.35 и 3.36 видно, что при переходе от амплитудно-структурной метрики к структурной в большей степени изменяются соотношения метрик в областях резких перепадов амплитуды сигнала. Стоит также отметить, что в областях низких амплитуд (как правило, в конце слова в области «придыхания») различного рода шумовые вклады в структуру ОТ-кластеров становятся значимыми, поэтому на таких участках метрики имеют большую величину.

В практических приложениях могут быть более полезны модифицированные методы анализа трендов и разладок. Например, для исключения случайных изменений структур и длительностей от одного периода ОТ к другому, следует вычислять метрики не смежных периодов, а разнесенных во времени друг от друга. В этом случае значения метрик в трендах станут

несколько выше по величине (так как в тренде дальше отстоящие друг от друга периоды ОТ будут сильнее отличаться друг от друга), а огибающая метрик станет плавнее. Также возможен вариант вычисления метрик при фиксированном положении одного периода ОТ и скользящем положении второго периода ОТ: в этом случае на результирующей картине с набором вычисленных метрик (положение отметки метрики в этом случае будет соответствовать положению скользящего периода ОТ) будут наблюдаться «провалы» на похожих по структуре звуках. Например, для уже рассматривавшегося слова «Сидорова» (произносится как [с’Идъръвъ]) при нахождении фиксированного кластера в одном из звуков [ъ] будут наблюдаться «провалы» метрик на двух других звуках [ъ].

Таким образом, амплитудно-структурные и структурные метрики, являясь наиболее прямым методом выявления изменений вокализованных звуков во временной области, могут быть использованы в алгоритмах автоматической сегментации вокализованных фрагментов речи в масштабе отдельных звуков.

3.7 Многопараметрические алгоритмы многоуровневой временной сегментации речевых сигналов

С учетом изложенных выше подразделов о частных алгоритмах различных уровней сегментации, в целом задача многоуровневой временной сегментации РС может быть разбита на несколько последовательных подзадач, каждая из которых представляет собой соответствующий уровень сегментации, например:

а. сегментация на активную речь и паузы+смычки (VAD-сегментация);

б. сегментация активной речи на шумные, вокализованные и взрывные звуки;

в. сегментация вокализованных звуков на отдельные периоды ОТ.

Примеры результата работы такого поэтапного по уровням алгоритма автоматической временной сегментации будут представлены ниже на эпюрах *«а»* рисунков 3.38...3.40. В данной реализации алгоритма одним из уровней сегментации является подалгоритм разделения активной речи на сегменты

«шумный / нешумный», поэтому взрывные звуки [т] и [к], имеющие шумовую структуру, относятся алгоритмом к классу шумных, а не взрывных.

Иным вариантом осуществления многоуровневой временной сегментации является организация параллельного вынесения частных решений на каждом интервале оценивания несколькими подалгоритмами сегментации. В данном случае сигнал во времени разбивается на окна равной длительности порядка:

$$n = \left\lfloor \log_2 \left(0.015 F_S \right) \right\rfloor^2, \tag{3.18}$$

где F_S – частота дискретизации речевого сигнала: для частоты дискретизации 22050 Гц n=256, для частоты 44100 Гц n=512 отсчетов.

В качестве подалгоритма могут выступать VAD-алгоритм, алгоритмы оценивания параметров средней мощности, частоты пересечений нуля, мел-частотных кепстральных коэффициентов MFCC и др. Решения, выдаваемые каждым отдельным подалгоритмом не обязательно должны покрывать весь перечень возможных типов сегментов. Например, VAD-алгоритм может выносить решения «речь / пауза», алгоритм оценки параметра средней мощности может выдавать решения «пауза / шумный / вокализованный слабый / вокализованный сильный». При этом для вынесения частных решений непосредственно используются статистические данные, получаемые на этапе исследования сигнальных особенностей звуков (раздел 2 «Исследование сигнальных особенностей звуков русской речи»).

Итоговое решение о типе сегмента на текущем интервале оценивания принимается на основе частных решений подалгоритмов – для этого обязательным этапом разработки алгоритма многоуровневой временной сегментации является этап обучения формированию итогового решения по частным. На данном этапе целесообразно также применять инструментарий, используемый в алгоритмах распознавания речи (построение нейронных сетей, марковских моделей).

На рисунке 3.37 для фонограммы слова «Трубка» показаны частные решения каждого из подалгоритмов на интервалах оценивания (под временной функцией РС) и итоговая автоматическая многоуровневая сегментация (над временной функцией).

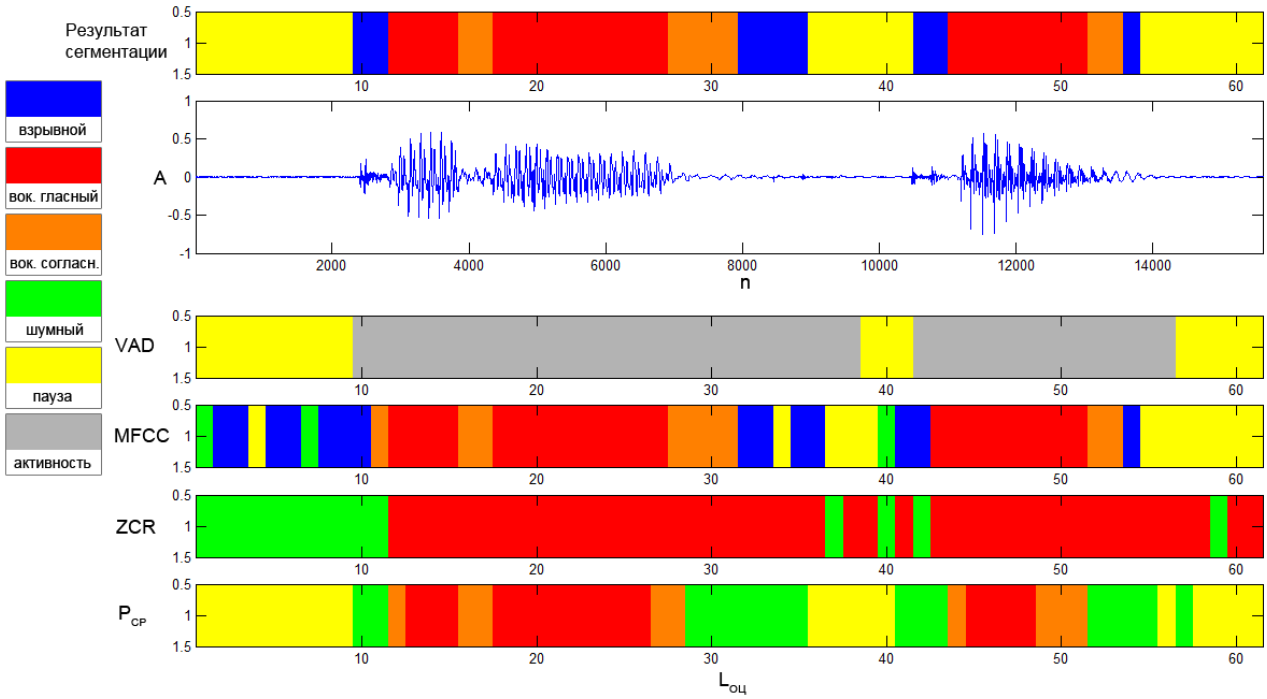


Рисунок 3.37 – Результаты работы частных подалгоритмов сегментации и итоговый результат автоматической МВС речевого сигнала слова «Трубка»

На выходе VAD-алгоритма возможны два решения: речевая активность; пауза / смычка.

На выходе подалгоритма, осуществляющего сегментацию на основе мел-частотных кепстральных коэффициентов (MFCC) возможны 5 решений: пауза / смычка; вокализованный гласный; вокализованный согласный; глухой шумный; глухой взрывной.

Для подалгоритма, основанного на параметре частоты пересечений нуля, возможны два решения: шумный; вокализованный (+ взрывной [п]).

Наконец, по параметру средней мощности звука на текущем интервале оценивания принимаются следующие частные решения: пауза / смычка; шумный звук; вокализованный сильный (вероятнее гласный); вокализованный слабый (вероятнее согласный).

Примеры результатов работы данного алгоритма представлены на эпюрах *«б»* рисунков 3.38...3.40 (цветовая маркировка соответствует легенде рисунка 3.37). В данном случае за счет реализации алгоритма с применением подалгоритма оценки мел-кепстральных коэффициентов становится возможным определение взрывных звуков [к], [к’], [т] непосредственно как взрывных, а не шумных. Так же становится возможным расширение алгоритма до приближений к пофонемной сегментации.

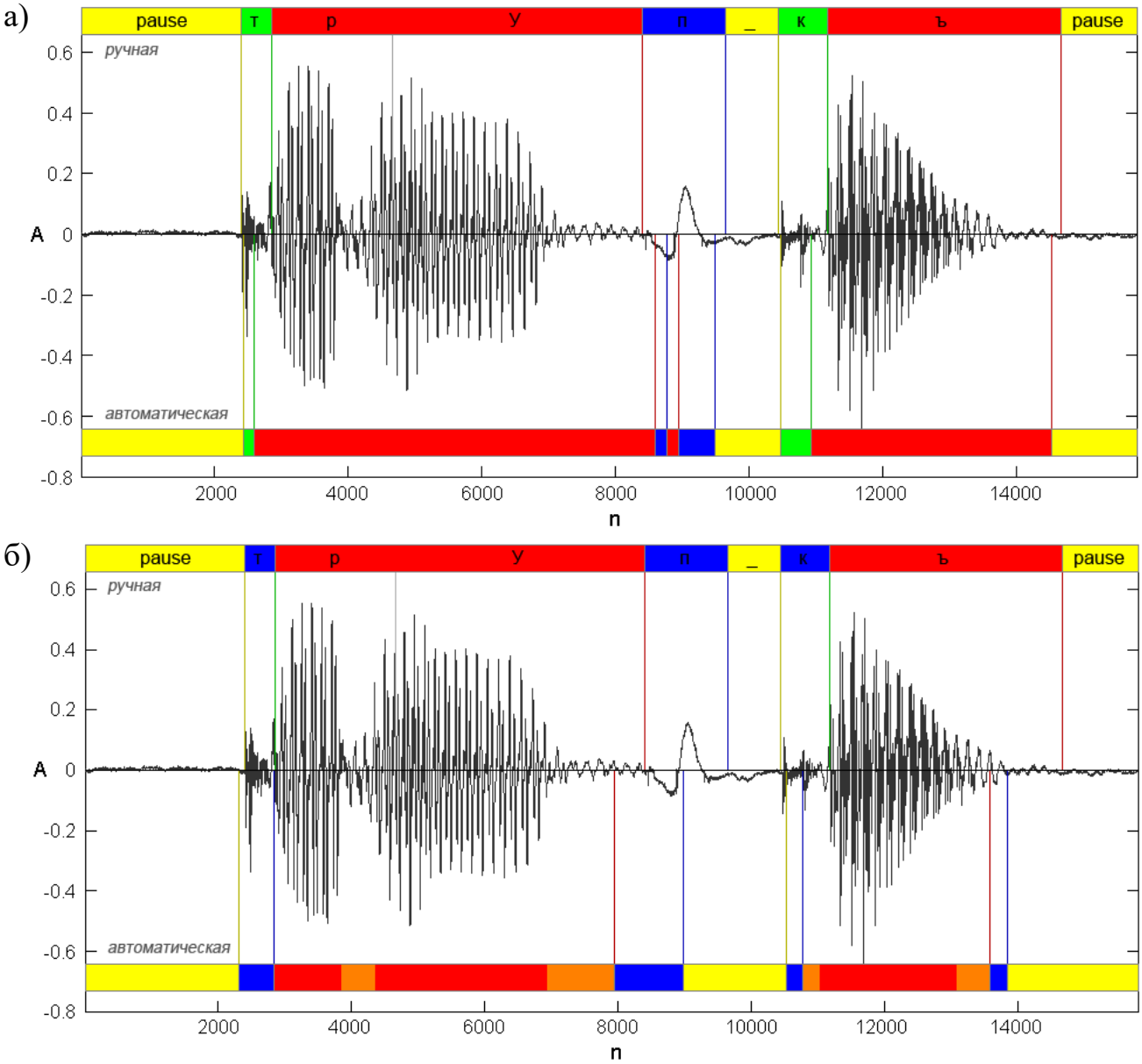


Рисунок 3.38 – Результаты работы алгоритмов МВС для фонограммы слова «Трубка», диктор мужчина: а) сегментация поэтапным алгоритмом; б) сегментация методом с параллельной работой подалгоритмов

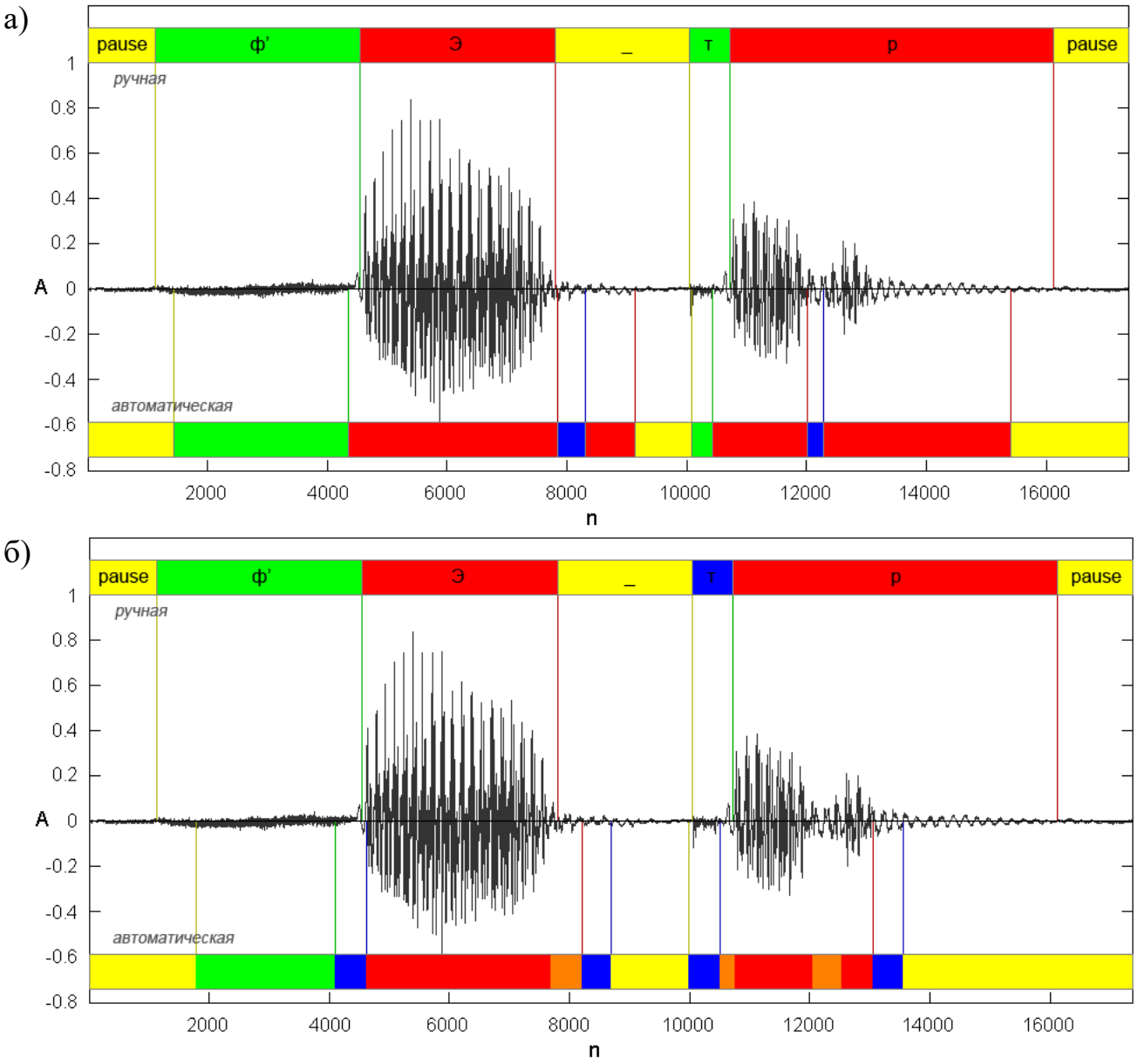


Рисунок 3.39 – Результаты работы алгоритмов МВС для фонограммы слова «Фетр», диктор мужчина: а) сегментация поэтапным алгоритмом; б) сегментация методом с параллельной работой подалгоритмов

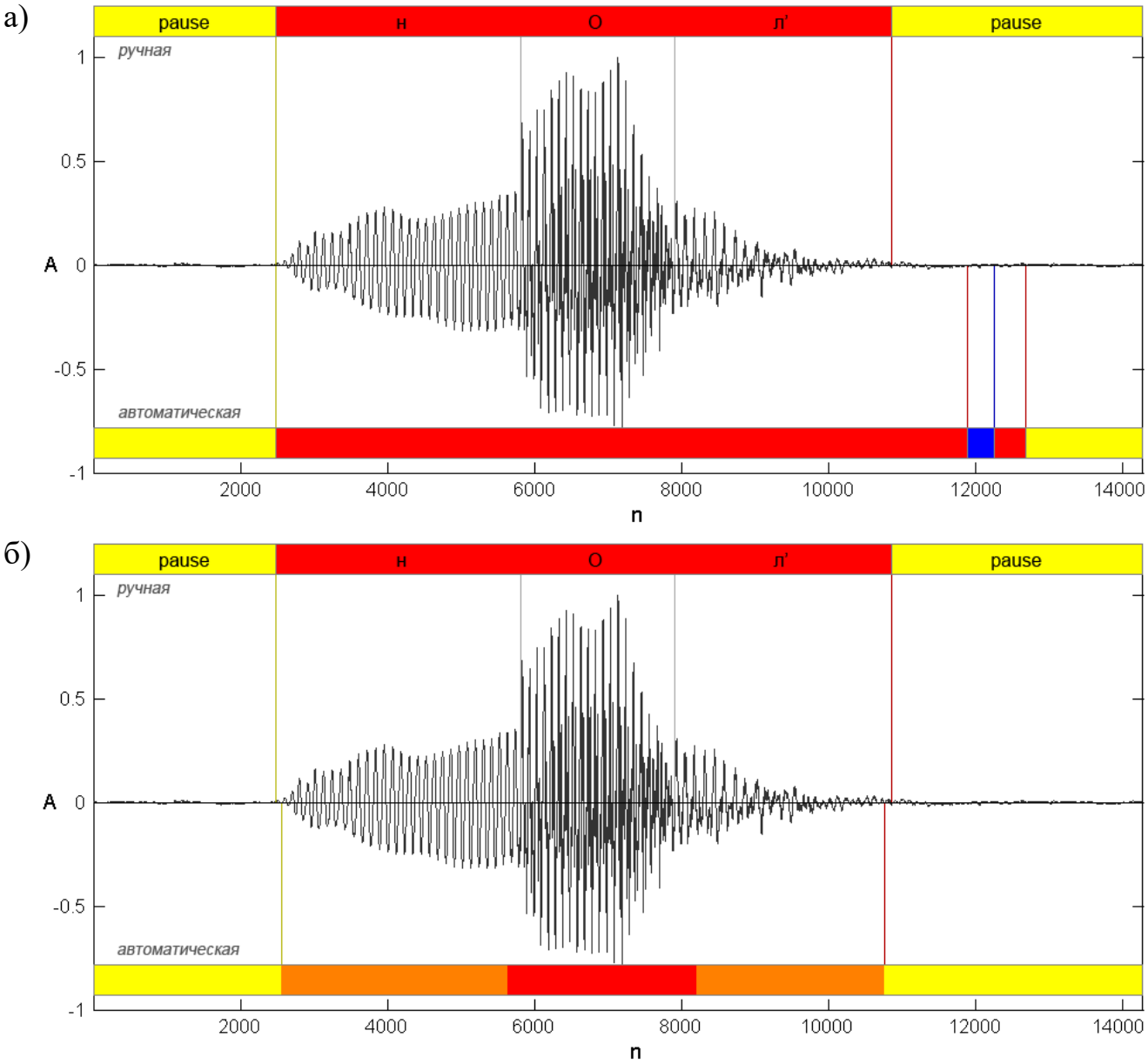


Рисунок 3.40 – Результаты работы алгоритмов МВС для фонограммы слова «Ноль», диктор женщина: а) сегментация поэтапным алгоритмом; б) сегментация методом с параллельной работой подалгоритмов

3.8 Основные выводы по разделу

При реализации процесса многоуровневой временной сегментации необходимый уровень может достигаться либо сразу за один этап, либо в несколько этапов с последовательным углублением уровня сегментации. Чаще всего сегментация производится в несколько этапов. При этом первым этапом, как правило, является VAD-сегментация, а на следующих этапах производится дополнительная сегментация активных участков речи.

Предложенный в рамках раздела метод сравнения эффективности однотипных (работа с одинаковым набором входных и выходных типов сегментов) алгоритмов сегментации позволяет количественно сравнить результаты работы алгоритмов сегментации вне зависимости от уровня сегментации, осуществляемой ими. Это дает возможность выбирать наиболее подходящие под конкретные условия конкретной задачи алгоритмы сегментации, а также устанавливать для них оптимальные настройки пороговых параметров.

Достаточно актуальным на практике является разработанный алгоритм выделения огибающей РС, формируемой непосредственно из точек локальных максимумов РС. В частности, на основе данного алгоритма разработан алгоритм сегментации по амплитудным разладкам РС, являющийся вспомогательным для других алгоритмов сегментации. Благодаря особенностям работы разработанного алгоритма выделения огибающей на вокализованных фрагментах РС, на его основе был разработан быстродействующий алгоритм ОТ-сегментации, выделяющий отдельные периоды колебаний голосовых связок диктора в точках амплитудных максимумов/минимумов.

В дополнение к алгоритму ОТ-сегментации, осуществляющему разбиение сигнала по моментам амплитудных максимумов, второй представленный алгоритм ОТ-сегментации является реализацией корреляционного метода и основывает выделение границ периодов ОТ на точках пересечения уровня нуля в направлении снизу-вверх низкочастотной составляющей РС.

Предложенная здесь же модификация алгоритма MFCC-параметризации позволяет более полно учитывать механизмы речеобразования и речевосприятия, что, как показали проведенные экспериментальные исследования, значительно увеличивает эффективность использования MFCC-коэффициентов в речевом приложении при низких ОСШ.

В рамках раздела рассмотрен вопрос увеличения эффективности энергетического VAD-алгоритма. Произведен сравнительный анализ исходного энергетического алгоритма, модифицированного алгоритма, а также стороннего

VAD-алгоритма [127] на нескольких фонограммах, записанных в разных условиях и разными дикторами.

Представленные в разделе другие алгоритмы поэтапной сегментации (сегментация «шумный/нешумный», «вокализованный/невокализованный») также обширно проверены на практике и успешно применены в речевом приложении изменения темпа произнесения речи (см. ниже подраздел 4.7 «Модификация произнесения речи»).

Наконец, показана состоятельность комплексного подхода к сегментации, основанного на независимой сегментации РС по разным параметрам. В реализации подхода важную роль играет накопленная статистическая информация по значениям параметров звуков русской речи (раздел 2 «Исследование сигнальных особенностей звуков русской речи»): по каждому применяемому параметру РС сегментируется на собственный набор типов сегментов, а затем по всем частным результатам сегментации формируется разбиение РС на требуемые по условию задачи типы сегментов.

4 ПРИЛОЖЕНИЯ РАЗРАБОТАННЫХ АЛГОРИТМОВ МНОГОУРОВНЕВОЙ ВРЕМЕННОЙ СЕГМЕНТАЦИИ РС

4.1 Функциональные алгоритмы обработки РС

В данном разделе рассматриваются некоторые функциональные алгоритмы – актуальные речевые задачи, – в работе которых важнейшую роль играет временная сегментация обрабатываемого РС. К таким задачам относятся:

1. Алгоритмы модификации темпа речи.
2. Алгоритмы сжатия речевого сигнала.
3. Алгоритмы командного управления (малый алфавит):
   - известный диктор;
   - произвольный диктор.
4. Использование голосового управления в контроле доступа (произнесение известного пароля или группы паролей известным диктором).
5. Идентификация и верификация диктора – производится сравнение опорной фонограммы по конкретному диктору с произнесенной им анализируемой фонограммой. При решении подобных задач необходимо сделать технологию менее субъективной. В случае алгоритмической верификации / идентификации оценка тождественности дикторов является не субъективной.
6. Выделение ключевых слов в потоке слитной речи.
7. Синтез речи
8. Шумоподавление в речевых сигналах

В таблице 4.1 приведены возможные комплексы технологических алгоритмов, применимых для решения перечисленных задач. Во второй половине раздела будет детально рассмотрен алгоритм модификации темпа речи в качестве примера использования результатов временной сегментации в речевых приложениях.

Таблица 4.1. Применение технологических алгоритмов обработки РС в составе функциональных

№ п/п	Функциональный алгоритм	Состав и особенности технологических алгоритмов
1	Сжатие речевых сигналов	1.1. – VAD 1.2. – Выделение тихих участков с учетом АЧХ слуха 1.3. – Равномерная по времени сегментация на фрагменты 5-30 мс 1.4. – VAD – В / Ш / Вз – Фонемы, последовательности вокализованных фонем 1.5. – В / не В – ОТ
2а	Распознавание команд	2.1. – VAD – Равномерная по времени сегментация – Параметризация 2.2. – VAD – В / Ш / Вз – ОТ – Фонемы – Параметризация фонем, периодов ОТ
2б	Выделение ключевых слов в потоке слитной речи	
2в	Распознавание слитной речи	
3	Идентификация и верификация диктора	3.1. – Сегментация на равные по времени фрагменты 5-30 мс 3.2. – VAD – В / Ш / Вз – ОТ 3.3. – VAD – В / Ш / Вз – ОТ – Фонемы, последовательности вокализованных фонем
4	Конкатенативный синтез речи	– VAD – В / Ш / Вз – ОТ – Фонемы, дифоны, трифоны
5	Шумоподавление	– VAD – В / Ш / Вз – ОТ
6	Модификация произнесения речи	– VAD – В / Ш / Вз – ОТ с точностью до фазы колебаний

4.2 Сжатие речевых сигналов

При оцифровывании стереозвука с частотой дискретизации 22 кГц и 16-тибитным квантованием для представления одной секунды записи требуется 704000 бит, или 88 кБ памяти. Применение последующего сжатия РС позволяет многократно уменьшить эти цифры и, соответственно, снизить требования к системам хранения и передачи речевой информации.

Учитывая, что паузы и смычки между звуками занимают большой процент времени в монологе, одним из вариантов сжатия является обнуление сегментов пауз, что становится возможным благодаря применению VAD-алгоритма. Возможность сжатия РС за счет пауз объясняется особенностями звуковосприятия человека, состоящими в терпимости слуха человека к удалению еле слышных звуков [36].

Основу для построения подобных методов сжатия звука с потерями дает существование порога слышимости: в фонограмме можно обнулить сэмплы, величина которых лежит ниже данного порога. Поскольку порог слышимости зависит от частоты, кодер должен основываться на спектре сжимаемого сигнала в каждый момент времени. Если сигнал для частоты f меньше порога слышимости этой частоты, то его следует отбросить.

Использование знаний психоакустики, а также применение различных методов сжатия для разных типов сегментов РС позволяет разрабатывать алгоритмы с высоким коэффициентом сжатия [18, 19].

4.3 Алгоритмы командного управления (малый алфавит)

При работе данного алгоритма диктором раздельно произносятся слова из перечня возможных команд, а самим алгоритмом выполняется автоматическая дешифрация произнесенной команды: из алфавита команд определяется команда, соответствующая произнесенному РС (рисунок 4.1).

Рисунок 4.1 – Функциональная схема блока распознавания команд

Поступающий на вход РС содержит команды, произносимые раздельно, поэтому не стоит задача выделения команд из слитного потока речи (блок 1 на рисунке 4.1). Важную роль в работе алгоритма играет точность временной сегментации произнесенной команды (блоки 2 и 3). В блоке 2 при этом в процессе сегментации или по ее результатам производится классификация сегментов на характерные типы: вокализованные, шумные, взрывные, паузы-смычки; формируются вектора параметров (ВП) сегментов (блок 4). Сегмент может быть описан либо стандартизованным набором параметров, не зависящим от типов фонем; либо формирование вектора параметров может производиться за два шага: определяется тип сегмента и затем вычисляется зависящий от типа ВП.

Получаемая последовательность сегментов, сопровождаемая соответствующими векторами параметров (блок 5), поступает на блок сравнения (блок 8) с эталонами из алфавита команд (блоки 6 и 7). Эталоны команд представляют собой такие же последовательности классифицированных на типы и параметризованных сегментов.

Результатом работы алгоритма является выбранная из алфавита команда, имеющая наибольшую меру сходства с произнесенной реализацией.

Точность, универсальность алгоритмов различается в зависимости от стоящих задач. В частности, алгоритм может работать с настройкой на определенного диктора либо являться дикторонезависимым. Точность определения команд определяется заложенными принципами формирования ВП фонем. Могут быть реализованы и сложные алгоритмы, показывающие точные результаты, и самые простые алгоритмы распознавания команд, работающие надежно для очень небольшого процента задач.

В общем случае распознавание команд в системах командного управления реализуется по последовательностям фонем или фрагментов РС равной длительности: последовательность, сформированная для текущей произнесенной команды, сравнивается с эталонными классифицированными и параметризованными последовательностями.

При сравнении команд могут учитываться различные характеристики РС. С добавлением каждой новой характеристики в работу алгоритма происходит его наращивание, углубление.

В простейших случаях (при крайне небольшом размере словаря и достаточном взаимном различии акустической формы команд) могут быть использованы даже такие параметры, как длительность команды, последовательность основных типов речевой активности, общее количество выделенных сегментов.

4.4 Идентификация и верификация диктора

Как показывает практика, индивидуальность голоса человека определяется двумя факторами: уникальными анатомическими особенностями речевого аппарата (длина тракта, размеры голосовых связок, расположение зубов и т.д.), и механизмом артикуляции, определяемым работой центральной нервной системы.

Анатомические особенности речевого аппарата определяют, в частности, геометрическую форму резонансных полостей, влияющую на характеристики формантных частот РС. Индивидуальная средняя частота основного тона, в свою очередь, зависит от длины, массы, силы натяжения голосовых связок.

Системы идентификации могут использовать статические и динамические признаки и их смеси:

1) Группа признаков, отражающая статические свойства речевого тракта (связаны с анатомическими особенностями человека):
   - средняя частота и дисперсия ОТ;
   - распределение периодов ОТ;
   - амплитудная модуляция (огибающая) периодов ОТ;
   - текущая частота ОТ;
   - соотношение длительностей звонких и шумных сегментов,
2) Группа признаков, отражающая динамические свойства (связаны с артикуляционной деятельностью),
   - просодические параметры речи (интонация, ритмика, громкость, ударения);
3) Смешанные группы признаков.

В зависимости от выбранного для решения задачи подхода, целесообразно применять следующие виды сегментации РС:

а. Разбиение высказывания на априорно заданное или произвольное количество сегментов равной длительности.
б. Разбиение высказывания на сегменты в моменты изменения какого-либо признака или набора признаков.
в. Разбиение высказывания на сегменты, соответствующие определенным фонологическим единицам (слова, фонемы, дифоны, трифоны и т.п.).

Для распознавания фонем, групп фонем и слов в основном применяются скрытые марковские модели, нейронные сети и их комбинации.

На рисунке 4.2 представлена обобщенная функциональная схема алгоритма идентификации диктора. Представленный подход допускает использование в качестве эталонов и анализируемых речевых последовательностей произвольные (не совпадающие) фразы.

Рисунок 4.2 – Обобщенная функциональная схема алгоритма верификации диктора

Описание блоков на схеме (рисунок 4.2):

1) Имеется база всех возможных идентифицируемых дикторов, для каждого из которых записана совокупность опорных фонограмм. В частном случае может использоваться по одной фонограмме на одного диктора, в общем случае объем записанного речевого материала для дикторов может быть разным. Следует также иметь в виду, что в имеющемся опорном материале может быть представлена не вся совокупность фонем русского языка, а лишь ее часть.
2) На вход алгоритма подается анализируемый РС, представляющий собой запись произнесенной неизвестным диктором фонограммы. Анализируемый материал также может быть разного объема.
3) И опорные, и анализируемые РС подвергаются временной сегментации фонемного уровня.

4) и 7) Сегментированные фрагменты РС классифицируются по основным типам фонем (вокализованные, взрывные, шумные, смычки). Производится их сопоставление с алфавитом фонем русского языка.
5) и 8) Для параметризации фонем могут применяться параметры, получаемые в разных областях рассмотрения сигнала: спектральные параметры, временные и т. д. Подробно вычисление вектора параметров фонем был рассмотрено выше.
6) и 9) Выполняется усреднение параметров нескольких реализаций одной фонемы. При усреднении предпочтителен (в особенности, при наличии большого числа реализаций одинаковых фонем) вариант построения гистограмм параметров – в данном случае становится возможным рассмотрение диапазона изменения определенных параметров для определенной фонемы, произнесенной определенным диктором. Еще одним вариантом проработки многократного повторения фонем в эталонной и/или выборочной фонограммах является их попарное сравнение – в данном случае по фонеме формируется обобщенное решение вида «из числа N реализаций фонемы x для текущего диктора характерно число фонем n».
10) Полученные наборы параметров для каждого диктора (в том числе анализируемая фонограмма рассматривается как произнесенная отдельным диктором) записываются в виде матрицы, структура которой показана на рисунке 4.3: *j* – номер фонемы в каталоге фонем (см. блоки 4 и 7 функциональной схемы); *i* – номер элемента в векторе параметров (как спектральных, так и не спектральных), в ячейках матрицы – усредненные значения соответствующего параметра. Некоторых фонем может не оказаться в конкретном записанном материале, поэтому не все фонемы языка могут быть представлены на оси *j*. Также необходимо отметить, что разные фонемы содержат разное количество полезной для верификации диктора информации.

Рисунок 4.3 – Вид матрицы средних значений параметров фонем для определенного диктора

Подалгоритмы, представляющие блоки (3), (4), (5), (6) рисунка 4.2 выполняются единожды заранее для всего банка опорных фонограмм. А подалгоритмы для блоков (3), (7), (8), (9) выполняются для каждой текущей исследуемой фонограммы.

Общий вид значений метрик (по формулам, предложенным выше), вычисленные для двух дикторов, показаны на рисунке 4.4. Для одного диктора (диктор *B* на рисунке), голос которого в большей степени соответствует анализируемой фонограмме, значения метрик оказываются в целом меньше, чем аналогичные значения, вычисленные в расчете на другого диктора (на рисунке – диктор *A*). Общее сравнение по всем фонемам производится на основе усреднения по фонемам. Для вынесения итогового решения об идентификации диктора также могут быть использованы мажоритарные («голосующие») алгоритмы.

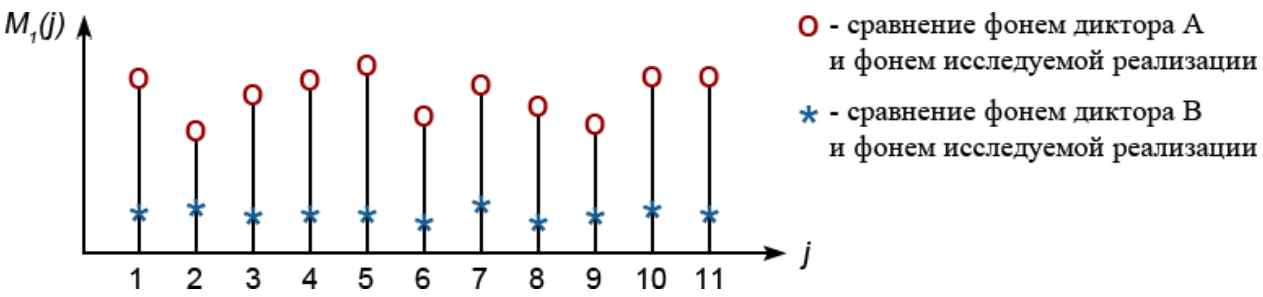


Рисунок 4.4 – Сравнение метрик параметров фонем при совпадении и несовпадении опорной и анализируемой фонограмм

4.5 Конкатенативный синтез речи

В конкатенативном синтезе для устранения артефактов звучания, связанных с различными уровнями сигнала окончания и начала смежных «склеиваемых» фрагментов, перед конкатенацией применяют оконное взвешивание для обеспечения плавного безразрывного перехода от одного речевого фрагмента к следующему (рисунок 4.5).

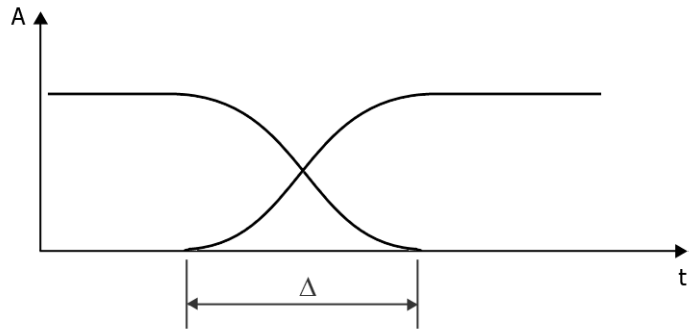


Рисунок 4.5 – Перекрытие взвешивающих окон при конкатенации фрагментов речи

Поэтому, для обеспечения возможности такого перекрытия перед нарезкой исходных фонограмм для добавления их фрагментов в базу конкатенативного синтеза необходимо изменить результаты их сегментации. Перед добавлением в базу границы всех фрагментов необходимо расширить по времени на величину $\Delta/2$ с каждой стороны сегмента (см. рисунок 4.5), тогда при конкатенации с перекрытием будет восстановлена истинная длительность фрагмента сигнала.

4.6 Шумоподавление

Применение эффективного VAD-алгоритма позволяет выявить участки фонограммы без речевой активности, состоящие исключительно из шумовой составляющей. Поэтому фрагменты пауз можно использовать для выявления статистических характеристик шума для использования их в настройках параметров очищающих фильтров. Учитывая тот факт, что паузы занимают значимую часть речи человека и часто появляются в РС, становится возможной адаптация к относительно медленно меняющимся шумам.

При рассмотрении вокализованных участков отчетливо отмечается значительная структурная схожесть рядом стоящих периодов ОТ – эту квазипериодичность можно использовать для повышения отношения сигнал/шум вокализованных фрагментов. Если каждый период ОТ обрабатывать с учетом формы рядом стоящих периодов, синфазно их складывая, то уровень полезной составляющей при таком сложении будет увеличиваться сильнее уровня шумов. На практике эта операция может осуществляться путем векторной интерполяции каждого периода ОТ с одним или несколькими смежными периодами (подробнее векторную интерполяцию периодов ОТ см. ниже в пункте 4.7.3 ниже). При этом за основу целесообразно брать длительность и энергию именно текущего периода, а не усредненные значения для всех интерполируемых векторов.

4.7 Модификация произнесения речи

4.7.1 Начальные сведения о модификации темпа речи

В текущем подразделе распространенная задача модификации речи представлена разработанным алгоритмом изменения темпа произнесения речи, в котором применены описанные в данной работе подалгоритмы сегментации, в частности VAD-алгоритм, двухэтапный алгоритм сегментации речевой активности на характерные типы, алгоритм ОТ-сегментации во временной области.

Операция изменения темпа позволяет так обработать РС, чтобы скорость произнесения изменялась в заданное количество раз, но при этом тембр (частота ОТ) голоса диктора оставалась без изменений. Можно привести следующие примеры использования изменения темпа речи:

- повышение комфортности восприятия речевой информации (настройка подходящей скорости воспроизведения при прослушивании аудиокниг, аудиоэкскурсий и т.п.);
- ведение оперативной стенографии (ручной перевод фонограммы в текстовую форму);
- быстрое прослушивание фонограмм, контролируемая перемотка;

– обеспечение необходимой скорости синтеза в синтезаторах речи;

– синхронизация аудио- и видеоканалов в видеороликах.

Выше в пункте 1.1.4 «Параметризация сегментов речевого сигнала» поднимались вопросы вариативности длительностей фонем в речи. Особенно важно отметить, что при разных темпах произнесения одной и той же фразы в естественной речи длительности разных типов фонем изменяются в разной степени.

В [133] на основе ряда экспериментов сделаны следующие выводы. При изменении темпа речи в большей степени изменяется длительность гласных. Относительная длительность согласных (отношение длительности согласных звуков к длительности гласных звуков) тем больше, чем выше темп произношения. При очень быстром темпе отдельные гласные могут исчезать, при больших замедлениях темпа длительность согласных звуков практически не меняется, а удлинение слогов происходит за счет гласных звуков. Длительность слов и слогов при изменении темпа речи также изменяется. Однако их относительная длительность (отношение длительности отдельных слогов и слов к общей длительности фразы) остается почти без изменений, то есть паузы между слогами и словами растягиваются и сжимаются примерно на ту же величину, что и активные участки речи.

4.7.2 Описание алгоритма модификации темпа произнесения речи

С учетом сказанного выше, является очевидным, что для изменения темпа речи (иными словами изменения длительности фонограммы при сохранении индивидуальных характеристик говорящего) необходимо, в первую очередь, произвести сегментацию РС, то есть разделить фонограмму на фрагменты пауз, вокализованных, шумных и взрывных участков речи. На рисунке 4.6 показана базовая последовательность операций при модификации темпа РС.

Рисунок 4.6 – Общая структура алгоритма модификации темпа речи

Блок 1. Для обработки исходная фонограмма должна быть представлена в цифровом виде, то есть должна быть дискретизирована по времени и квантована по уровню. В этом же блоке могут производиться и другие операции по подготовке фонограммы для модификации ее темпа, например, нормализация и устранение смещения по постоянному току.

Блок 2. Для разделения РС на фрагменты активности говорящего и пауз между ними применяется VAD-алгоритм.

Блок 3. Блок разделения активных участков речи на характерные типы является сравнительно сложным, и в нем можно выделить подблоки, назначение и взаимодействие которых определяется выбранными методами сегментации и методами модификации характерных типов. На рисунке 4.7 показан пример возможной организации подблоков.

Блок 4. Оператор вводит интегральный коэффициент растяжения K, ожидая получить на выходе алгоритма фонограмму, длительность звучания которой в K раз больше длительности звучания исходной.

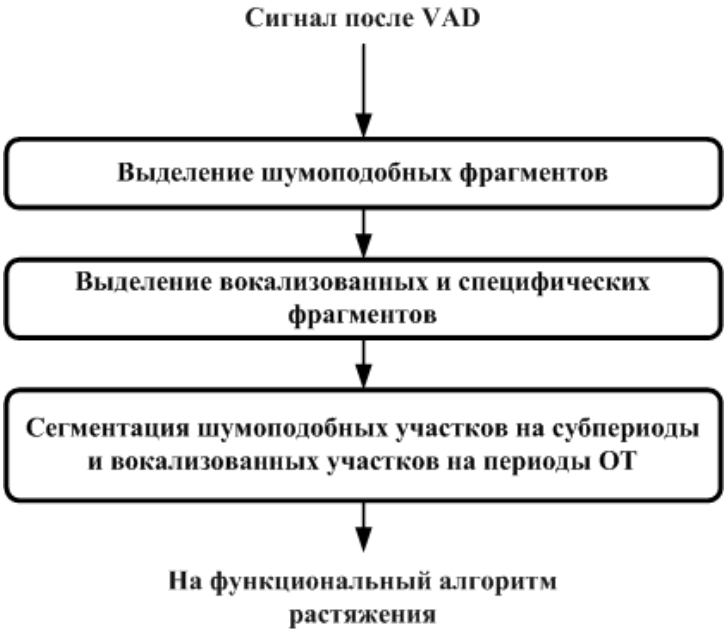


Рисунок 4.7 – Пример подструктуры блока сегментации активных участков речи

Блок 5. По результатам сегментации фонограммы определяется темп речи диктора исходной фонограммы. По заданному оператором интегральному коэффициенту растяжения K и с учетом статистики присутствия различных звуков в речи вычисляются парциальные коэффициенты $K_{вок}$, $K_{шум}$, $K_{пауз}$ для непосредственной модификации соответственно вокализованных, шумных фрагментов и фрагментов пауз. Специфические участки (взрывные, ударные, переходные) имеют сложную структуру и при этом непродолжительны по длительности. Как показано в работах [134, 135], такие участки РС при изменении темпа можно оставлять без изменений, поэтому для них парциальный коэффициент растяжения $K_{вз}=1$.

Блок 6. По результатам сегментации фонограммы применяются выбранные частные для типов фрагментов алгоритмы растяжения с рассчитанными парциальными коэффициентами в качестве входных параметров, и тем самым производится модификация исходной фонограммы. В результате, средний коэффициент изменения темпа речи в фонограмме соответствует заданному интегральному коэффициенту K.

Блок 7. В зависимости от стоящей задачи, выполняются озвучивание, и/или сохранение в файл модифицированного РС, и/или передача его в другие речевые приложения для дальнейшей обработки.

На рисунке 4.8 приведена структура выполненной программной реализации разработанного алгоритма.

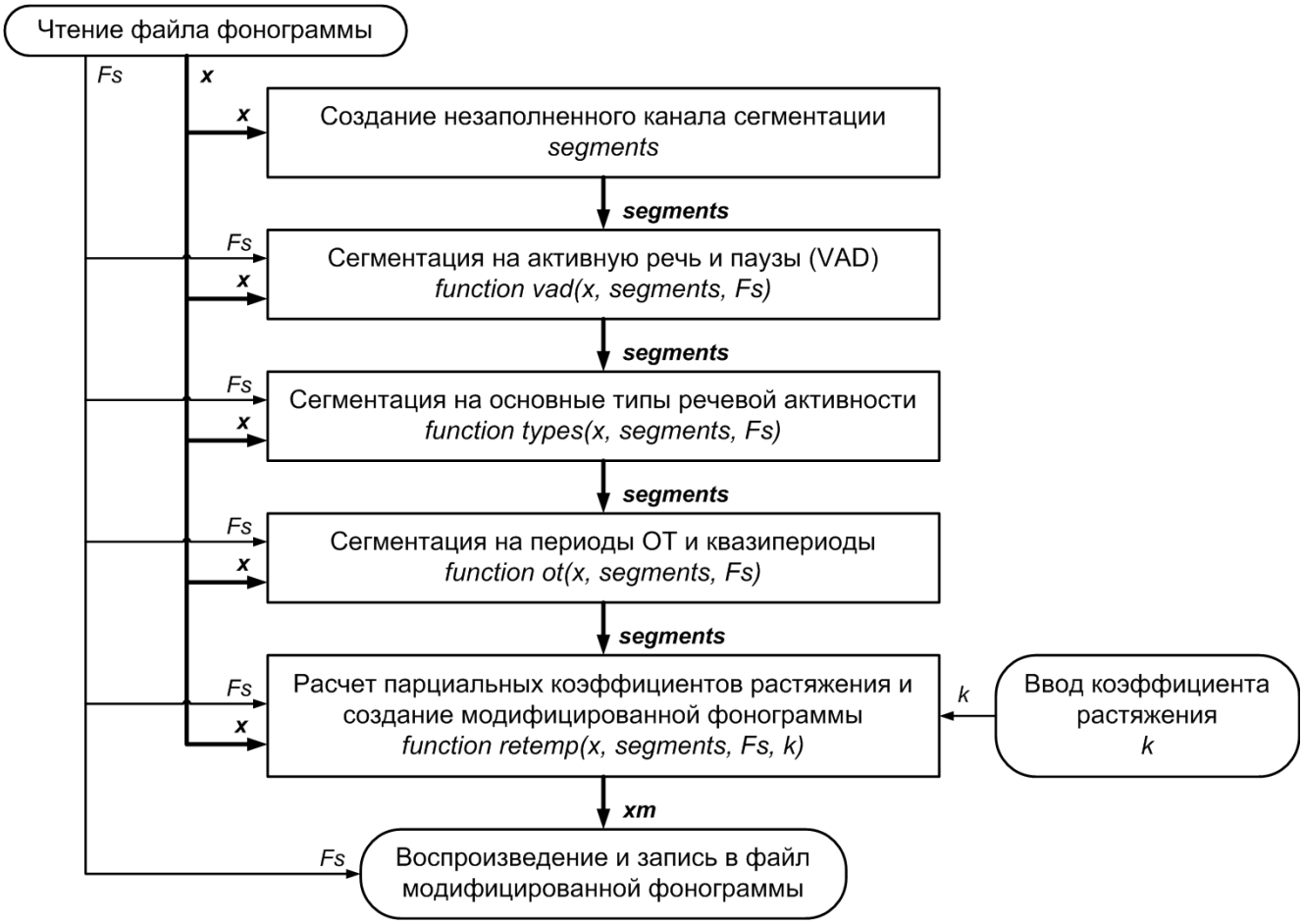


Рисунок 4.8 – Структурная схема программы модификации темпа произнесения речевых фонограмм

4.7.3 Изменение темпа произнесения для пауз и различных типов фонем

Как было сказано ранее, наиболее сильной модификации следует подвергать вокализованные фрагменты речи и участки пауз между словами и частями слов. Шумные же звуки необходимо изменять в значительно меньшей степени, а взрывные и вовсе оставлять без изменений.

По введенному оператором интегральному коэффициенту растяжения *K*, алгоритмом модификации темпа речи должны быть подобраны парциальные коэффициенты для обеспечения итогового изменения длительности фонограммы в близкое к заданному *K* количество раз. Такой расчет должен осуществляться по статистическим данным о средней доле времени, приходящейся на каждый тип звука в естественной речи. Для более точного соответствия длительности модифицированной фонограммы задаваемому коэффициенту *K*, следует произвести оценку темпа произнесения исходной фонограммы. Это можно сделать, обладая упомянутыми выше статистическими данными, после сегментации репрезентативного участка исходной фонограммы и вычисления соотношения длительностей звуков разных типов в нем. Следует также отметить, что в монологе темп произнесения может меняться в зависимости от эмоциональной окраски текущего контекста, поэтому при обработке больших по длительности фонограмм и при обработке в реальном времени возникает необходимость следить за текущим темпом речи говорящего.

Для изменения темпа за счет вокализованных участков необходимо изменять в них количество периодов ОТ. При этом при растяжении производится добавление, а при сжатии – удаление отдельных периодов. Для лучшего качества звучания модифицированной фонограммы применяется векторная интерполяция двух периодов ОТ исходной фонограммы (рисунок 4.9). При растяжении между двумя исходными периодами ОТ вставляются один или несколько (в зависимости от коэффициента растяжения) синтезированных периодов. При сжатии же два или более периодов исходной фонограммы замещаются одним периодом – результатом их интерполяции. При векторной интерполяции учитывается возможность несовпадения длительности исходных периодов. А также, в случае синтеза из двух исходных нескольких векторов, учитывается условие необходимости плавного изменения длительности и величины синтезируемых векторов от соответствующих значений первого исходного до соответствующих значений второго исходного векторов [126]. Примеры результатов интерполяции показаны на рисунке 4.9.

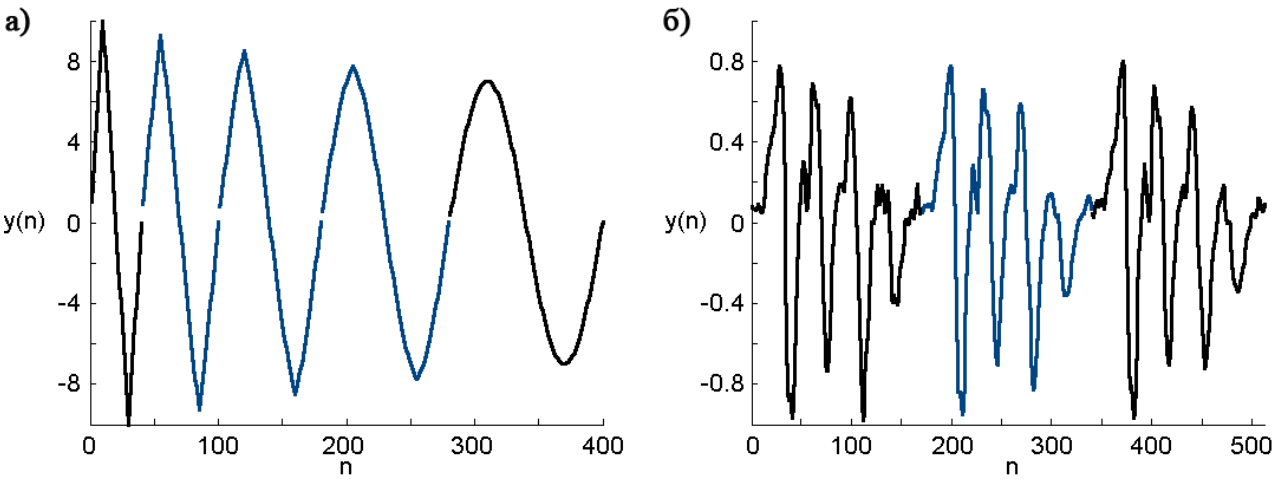


Рисунок 4.9 – Визуализация интерполяции периодов ОТ: а) по двум исходным исскусственным векторам ОТ-колебаний синтезируется три переходных вектора; б) синтезированный период ОТ (в центре) реального РС

Для изменения длительности за счет шумных звуков можно предложить несколько подходов. Действенным оказывается метод, применяемый для модификации вокализованных фрагментов: добавляемый участок является переходным между предыдущим и либо смежным следующим, либо случайно выбранным участком текущего шумного звука. Однако для экономии машинного времени возможен вариант копирования случайно выбранных участков такого звука. При замедлении темпа из шумных сегментов удаляются участки необходимой длительности. Для обеспечения неразрывности получаемого сигнала участки должны добавляться и удаляться в точках пересечения сигналом нуля в определенном направлении (пересечение сверху вниз или снизу вверх). Поэтому при модификации шумного участка следует его предварительно сегментировать на субпериоды: интервалы между двумя переходами через нулевой уровень в выбранном направлении.

Специфические участки (взрывные, ударные, переходные), как уже говорилось, целесообразно оставлять без изменений.

В случае участков пауз, которые являются маломощными и еле слышимыми, можно, как и в случае шумных фрагментов, предложить несколько подходов. Во-первых, всю паузу можно рассматривать как один сегмент, который

либо частично копируется при замедлении темпа произнесения, либо не полностью воспроизводится при ускорении темпа (опять же, желательно производить копирование и удаление в точках пересечения нуля). Если же позволяют вычислительные ресурсы аппаратуры для повышения качества звучания возможно применение интерполяции участков паузы.

Пример участка фонограммы, полученной на выходе разработанного алгоритма модификации темпа произнесения речи, показан на рисунке 4.10.

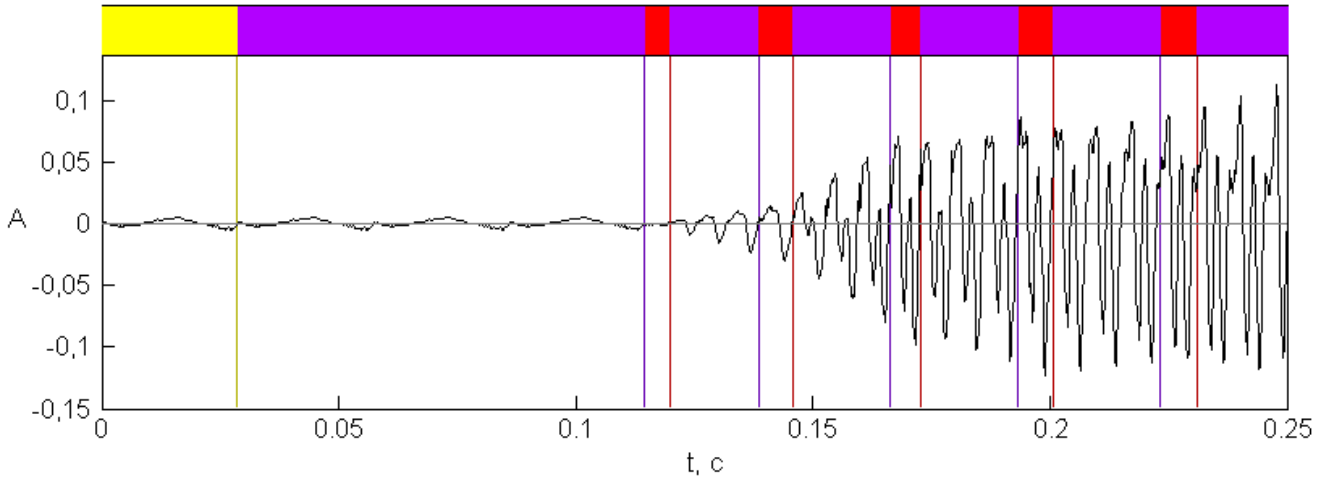


Рисунок 4.10 – Участок фонограммы с замедленным темпом произнесения (желтым цветом показан сегмент «пауза», красным цветом – «вокализованный» сегмент, фиолетовым цветом – синтезированные алгоритмом фрагменты РС)

4.7.4 Анализ эффективности алгоритма модификации темпа речи

Для анализа эффективности разработанного алгоритма модификации темпа речи произведено его сравнение методом экспертных оценок с рядом существующих программных средств (таблица 4.2), решающих аналогичную задачу.

Таблица 4.2 – Сторонние программные решения задачи модификации темпа речи

Название	Версия	Дата релиза
Audipo	2.3.2	5 августа 2016
Audio Speed Changer Pro	1.5.5	27 сентября 2012
AIMP	4.02	30 мая 2016
Sony Sound Forge Pro	11.0	13 января 2015

При проведении опроса членам экспертной группы было предложено выполнить ранжирование набора фонограмм «в порядке от самой качественной до самой некачественной, принимая во внимание разборчивость слов и комфортность восприятия записи». Всего для анализа было подобрано три фонограммы (таблица 4.3), каждая из которых была модифицирована сравниваемыми алгоритмами с четырьмя коэффициентами ускорения/замедления темпа: ускорение в 1,5 и 2 раза; замедление в 2 и в 3 раза. Таким образом, каждому эксперту всего необходимо было выполнить 12 этапов ранжирования пяти объектов, для каждого эксперта на каждом этапе ранжирования фонограммы предлагались в квазислучайном порядке.

Таблица 4.3 – Исходные тестовые фонограммы

Номер	**Описание**	**Параметры**	**Длительность**
1	Женский голос: ведущая новостей, быстрый темп речи	7 бит, 8 кГц	4 с
2	Мужской голос: ведущий новостей	8 бит, 22 кГц	6 с
3	Аудиокнига, мужской и женский голос, артистичная артикуляция, быстрый темп речи мужского голоса	16 бит, 22 кГц	7 с

По полученным от экспертов оценкам были сформированы две сводные таблицы: результаты ранжирования в режиме замедления (таблица 4.4) и ускорения (таблица 4.5) темпа речи, – в обоих случаях разработанный алгоритм занимает лидирующую позицию.

Таблица 4.4 – Ранги, присвоенные алгоритмам в режиме замедления темпа речи

Объект / Эксперт	$Э_1$	$Э_2$	$Э_3$	$Э_4$	$Э_5$	$Э_6$	$Э_7$	$Э_8$	$Э_9$	r_i
Разработанный алгоритм	1	1	1	1,5	1	1	1	1	1	9,5
Sony Sound Forge	2	2	2	1,5	2	2	2	2	2	17,5
Audio Speed Changer Pro	4	3,5	3	4	3	3	4	3	4	31,5
Audipo	5	3,5	4	3	4	4	3	4	3	33,5
AIMP	3	5	5	5	5	5	5	5	5	43

Таблица 4.5 – Ранги, присвоенным алгоритмам в режиме ускорения темпа речи

Объект / Эксперт	**Э$_1$**	**Э$_2$**	**Э$_3$**	**Э$_4$**	**Э$_5$**	**Э$_6$**	**Э$_7$**	**Э$_8$**	**Э$_9$**	**r$_i$**
Разработанный алгоритм	1	1	1,5	1	1	1,5	1	1	1	10
Sony Sound Forge	3,5	3	1,5	3	3	3	3	2	3,5	25,5
Audio Speed Changer Pro	3,5	2	4	2	2	4	2	3	3,5	26
Audipo	2	5	3	5	4	5	4	4	2	34
AIMP	5	4	5	4	5	1,5	5	5	5	39,5

Для определения тесноты связи между полученными ранжировками необходимо произвести количественную оценку степени согласованности мнений экспертов. При установлении статистической связи между оценками нескольких (более двух) объектов ранжирования применяется дисперсионный коэффициент множественной конкордации Кендалла [136]. При наличии в ранжировках связанных рангов данный коэффициент W вычисляется по формуле:

$$W = \frac{12 \cdot \sum_{i=1}^{m} \left(r_i - \bar{r} \right)^2}{d^2 \cdot (m^3 - m) - d \cdot \sum_{s=1}^{d} T_s}, \tag{4.1}$$

где d – количество экспертов, m – количество объектов ранжирования, r_i – сумма рангов i-го объекта (см. крайний правый столбец в таблицах 4.4 и 4.5), $\bar{r}$ – оценка математического ожидания значений r_i, T_s – показатель связанных рангов в s-й ранжировке:

$$T_s = \sum_{k=1}^{H_s} \left(h_k^3 - h_k \right), \tag{4.2}$$

где H_s – число групп равных рангов в s-й ранжировке, h_k – количество равных рангов в k-й группе ранжировки.

Коэффициент конкордации W (4.1) может принимать значения от нуля до единицы, при значениях $W > 0{,}5$ оценки экспертов считаются в достаточной мере согласованными [137]. В режиме замедления темпа речи рассчитанный коэффициент конкордации составляет $W = 0{,}89$, в режиме ускорения – $W = 0{,}63$, что позволяет сделать вывод о достаточной согласованности мнений экспертов. В режиме ускорения темпа различия между фонограммами, полученными разными алгоритмами, достаточно сложно выявить на слух, чем можно объяснить значительное уменьшение степени согласованности мнений экспертов по сравнению с режимом замедления.

4.8 Основные выводы по разделу

В разделе проработаны особенности реализации основных функциональных алгоритмов обработки РС в аспекте использования результатов временной сегментации и параметризации.

Показано, что сегментация является важнейшим этапом в реализации данных функциональных алгоритмов. В то же время, в зависимости от требований к быстродействию, качеству обработки исходного РС, для одних и тех же задач могут успешно применяться различные уровни сегментации и различная точность формирования ВП. В разделе описаны возможные подходы к реализации функциональных алгоритмов с применением тех или иных технологических алгоритмов обработки РС.

Описан разработанный, реализованный программно на языке MATLAB и успешно апробированный алгоритм модификации темпа произнесения речи. Работа данного алгоритма основана на многоуровневой временной сегментации РС до характерных типов звуков и отдельных периодов ОТ. Для различных типов звуков применяются разные парциальные коэффициенты и подалогритмы модификации РС, что позволяет получать комфортно воспринимаемые на слух модифицированные фонограммы.

За счет использования сегментации речевого сигнала разработанный алгоритм лишен ряда источников артефактов звучания, характерных для

существующего основного подхода к решению задачи модификации темпа речи, основанного на принципах работы вокодера. В частности, в разработанном алгоритме отсутствует эффект реверберации, свойственный вокодерным методам в режиме замедления темпа, а также чрезмерное редуцирование отдельных коротких звуков, свойственное вокодерным методам при значительном (двукратном) ускорении темпа речи. В результате, при сравнении группой экспертов выходных фонограмм ряда алгоритмов модификации темпа речи, разработанный в рамках диссертационной работы алгоритм занял лидирующую позицию как в режиме замедления, так и в режиме ускорения темпа.

ЗАКЛЮЧЕНИЕ

Основные результаты диссертационной работы можно обозначить следующими положениями:

1. Охарактеризован круг задач, решаемых с применением временной сегментации речевых сигналов, представлена соответствующая классификация применяемых в речевых приложениях уровней сегментации. Произведен обзор существующих общих подходов к автоматической сегментации и методов решения частных наиболее распространенных задач сегментации.
2. Выполнен анализ сигнальных особенностей звуков на примере русского языка. Предложен перечень аллофонов, характеризующий основные варианты произнесения русских фонем: данный перечень представляется достаточно кратким с точки зрения всего многообразия существенных и несущественных аллофонов и достаточно полным с точки зрения описания звучания речевых сигналов. Представлена таксономия существенных аллофонов русского языка, полученная на основе исследования их сигнальных особенностей
3. Создан программный комплекс, объединяющий файловое хранилище цифровых фонограмм, хранилище базы данных, функционал для обработки речевых сигналов, текстов, систематизации данных. С помощью созданного программного комплекса, в частности, осуществляется автоматизированное транскрибирование русских слов, ручная фонемная сегментация записанных фонограмм, вычисление сигнальных параметров звуков, формирование выборок вычисленных значений параметров для произвольных групп аллофонов / типов речевой активности.
4. Разработаны и апробированы как вспомогательные для сегментации методы, так и частные алгоритмы сегментации речевого сигнала на различные уровни: речь/пауза, вокализованный/взрывной/шумный, сегментация на периоды основного тона.

5. Предложен универсальный метод численной оценки эффективности произвольного алгоритма временной сегментации речевого сигнала. На основе данного метода выполнен анализ эффективности разработанной модификации энергетического VAD-алгоритма.
6. Разработан алгоритм модификации в широком диапазоне темпа произнесения речевых фонограмм, основанный на использовании результатов многоуровневой временной сегментации и показывающий высокую эффективность в сравнении с существующими аналогами.

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

АКФ – автокорреляционная функция

В – вокализованный сегмент

Вз – взрывной сегмент

ВКФ – взаимокорреляционная функция

ВП – вектор параметров

КФ – корреляционная функция

МИР – минимум информационного рассогласования

МФ – модулирующая функция

НККФ – нормированная кросс-корреляционная функция

ОТ – основной тон; ОТ-сегментация – сегментация на отдельные периоды ОТ

ПАМ – психоакустическая модель

СКО – среднеквадратическое отклонение

ШФК – широкий фонетический класс

РС – речевой сигнал

ФВЧ – фильтр верхних частот

ФНЧ – фильтр низких частот

Ш – шумный сегмент

ATH – Absolute Threshold of Hearing, абсолютный порог слышимости

DTX – Discontinuous Transmission, прерывистая передача

GSM – Global System for Mobile Communications, глобальная система мобильной связи

LPC – Linear Predictive Coding, кодирование с линейным предсказанием

MFCC – Mel Frequency Cepstral Coefficients, мел-частотные кепстральные коэффициенты

MSF – Magnitude Sum Function, функция суммы модулей

STM – Spectral Transition Measure, мера спектрального перехода

VAD – Voice Activity Detection, определение границ активной речи

ZCR – Zero Crossing Rate, частота пересечений нуля

СПИСОК ЛИТЕРАТУРЫ

1 **Кипяткова, И. С.** Автоматическая обработка разговорной русской речи : монография / И. С. Кипяткова, А. Л. Ронжин, А. А. Карпов. СПИИРАН. – СПб. : ГУАП, 2013. – 314 с.

2 **Бердников, О. М.** Модель пофонемного розпізнавання мови на основі акустичних параметрів смугового вокодеру / О. М. Бердников, К. Ю. Богуш // Збірник наукових праць / Військовий інститут телекомунікацій та інформатизації Національного технічного університету України «Київський політехнічний інститут». – Випуск № 2. – Київ : ВІТІ НТУУ «КПІ», 2010. – С. 11–18.

3 **Galunov, V. I.** From artificial intelligence to smart environment – on the problem of speech recognition / V. I. Galunov, N. G. Kouznetsov, A. N. Soloviev // International workshop «Speech and Computer» Proceedings / SPb, Russia. – 2004. – P. 405–410.

4 **Rambabu, D.** Speech Recognition of Industrial Robot / D. Rambabu, R. Naga Raju, B. Venkatesh // International journal of computational mathematical ideas. – 2011. – Vol. 3. – No. 2. – P. 92–98.

5 **Juang, B. H.** Automatic speech recognition – a brief history of the technology development / B. H. Juang, L. R. Rabiner // Elsevier Encyclopedia of Language and Linguistics. – Second edition. – 2005. – P. 806–819.

6 **Фланаган, Д. Л.** Анализ, синтез и восприятие речи / Д. Л. Фланаган. – М. : Связь, 1968. – 396 с.

7 **Рабинер, Л. Р.** Теория и применение цифровой обработки сигналов / Л. Р. Рабинер, Б. Голд. – М. : Мир, 1978. – 848 с.

8 **Рабинер, Л. Р.** Цифровая обработка речевых сигналов : [пер. с англ.] / Л. Р. Рабинер, Р. В. Шафер; под ред. М. В. Назарова и Ю. Н. Прохорова. – М. : Радио и связь, 1981. – 496 с.

9 Применение цифровой обработки сигналов : [пер. с англ.] / под ред. Э. Оппенгейма. – М. : Мир, 1980. – 552 с.

10 **Маркел, Дж. Д.** Линейное предсказание речи : [пер. с англ.] / Дж. Д. Маркел, А. Х. Грэй; под ред. Ю. Н. Прохорова и В. С. Звездина. – М. : Связь, 1980. – 308 с.

11 **Томчук, К. К.** Высококачественный алгоритм модификации темпа произнесения речи: разработка и апробация / К. К. Томчук, А. Ю. Зилинберг, Ю. А. Корнеев // Международная научная конференция «Системы и модели в информационном мире (СМИ-2009)»: материалы конференции / Таганрог : ТТИ ЮФУ (ТРТУ), 2009. – С. 80–91.

12 **Melin, H.** On Word Boundary Detection in Digit-Based Speaker Verification / H. Melin // Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C) / Avignon, France. – 1998. – P. 46–49.

13 **Сорокин, В. Н.** Сегментация и распознавание гласных / В. Н. Сорокин, А. И. Цыплихин // Информационные процессы. – 2004. – Т. 4. – № 2. – С. 202–220.

14 **Клименко, Н. С.** Разработка структуры текстонезависимой системы идентификации диктора / Н. С. Клименко // Искусственный интеллект. – 2012. – № 4. – С. 161–171.

15 **Бурибаева, А. К.** Сегментация и дифонное распознавание речевых сигналов / А.К. Бурибаева, Г.В. Дорохина, А.В. Ниценко, В.Ю.Шелепов // Труды СПИИРАН. – 2013. – № 8(31). – С. 20–42.

16 **Вишнякова, О. А.** Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования / О. А. Вишнякова, Д. Н. Лавров // Математические структуры и моделирование. / Омск : Ом. гос. ун-т, 2011. – Вып. 23. – С. 43–48.

17 **Petrushin, V. A.** Pitch-Synchronous Speech Signal Segmentation and Its Applications / V. A. Petrushin // Text, Speech and Dialogue. – 2003. – Vol. 2807. – pp. 321–326.

18 **Kanade, J. B.** A literature survey on psychoacoustic models and wavelets in audio compression / Jagadeesh B. Kanade, Dr. Sivakumar B. // International Journal of

Advanced Research in Electronics and Communication Engineering (IJARECE). – 2014. – Vol. 3. – Issue 1. – P. 1–7.

19 **Bardenhagen, S. T.** Low bit rate speech compression using hidden markov modeling / S. T. Bardenhagen, K. L. Brown, R. D. Braun // Proceedings of IEEE Military Communications Conference (MILCOM), Monterey, USA – 1997. – Vol. 1. – P. 507–511.

20 **Романенко, В. О.** Эмоциональные характеристики вокальной речи и их связь с акустическими параметрами / В. О. Романенко // Общество. Среда. Развитие (Terra Humana). – 2011. – № 3. – С. 124–127.

21 **Ipsic, I.** Speech technologies / I. Ipsic. – Rijeka, Croatia : InTech, 2011. – 432 p.

22 **Зилинберг, А. Ю.** Анализ характеристик импульсных помех в тракте передачи речевых сигналов / А. Ю. Зилинберг, Ю. А. Корнеев, К. К. Томчук // Сборник докладов Научной сессии ГУАП / СПб. : ГУАП, 2011. – Ч. 2. – С. 19–20.

23 **Зилинберг, А. Ю.** Разработка алгоритмов подавления импульсных помех в трактах передачи речевых сигналов / А. Ю. Зилинберг, Ю. А. Корнеев, К. К. Томчук // Сборник докладов Научной сессии ГУАП / СПб. : ГУАП, 2011. – Ч. 2. – С. 20–23.

24 **Ganapathiraju, A.** Syllable-Based Large Vocabulary Continuous Speech Recognition / A. Ganapathiraju, J. Hamaker, J. Picone, G. R. Doddington, M. Ordowski // IEEE Transactions on Speech and Audio Processing. – 2001. – Vol. 9. – No. 4. – P. 358–366.

25 **Wilpon, J. G.** An investigation on the use of acoustic sub-word units for automatic speech recognition / J. G. Wilpon, B. H. Juang, L. R. Rabiner // Proc. of International Conference Acoustic, Speech, and Signal Processing / Dallas, TX, USA. – 1987. – P. 821–824.

26 **Prasad, V. K.** Automatic segmentation of continuous speech using minimum phase group delay functions / V. K. Prasad, T. Nagarajan, H. A. Murthy // Speech Communication. – 2004. – Vol. 42. – P. 429–446.

27 **Rabiner, L. R.** A bootstrapping training technique for obtaining demisyllable reference patterns / L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, T. M. Zampini //

The Journal of the Acoustical Society of America (JASA). – 1982. – Vol. 71. – No. 6. – P. 1588–1595.

28 **Мещеряков, Р. В.** Алгоритмы оценки автоматической сегментации речевого сигнала / Р. В. Мещеряков, А. А. Конев // Информатика и системы управления. – 2012. – № 1 (31). – С. 195–206.

29 **Greenberg, S.** Strategies for automatic multi-tier annotation of spoken language corpora / S. Greenberg // Proc. of 8th European Conference on Speech Communication and Technology, EUROSPEECH-2003 / Geneva, Switzerland. – 2003. – P. 45–48.

30 **Бухаева, О. Д.** К сегментации речевого потока в русском языке в аспекте порождения речи / О. Д. Бухаева // Ученые записки Забайкальского государственного гуманитарно-педагогического университета им. Н. Г. Чернышевского (серия «Филология, история, востоковедение») / Чита : ЗабГГПУ, 2012. – № 2 (43). – С. 19–23.

31 **Киселев, В. В.** Система пофонемного автоматического распознавания команд русской речи / В. В. Киселев, И. Б. Тампель, М. Ю. Татарникова, Ю. Ю. Хохлов // Труды международной конференции «Диалог–2007»: Компьютерная лингвистика и интеллектуальные технологии / М. : Наука, 2007. – С. 236–241.

32 **Мазуренко, И. Л.** О сокращении перебора в словаре речевых команд в составе системы распознавания речи / И. Л. Мазуренко // Интеллектуальные системы / М. : МГУ, 1997. – Т. 2. – Вып. 1–4. – С. 135–148.

33 **Федоров, В. М.** Сегментация сигналов на основе дискретного вейвлет-преобразования / В. М. Федоров, П. Ю. Юрков // Информационное противодействие угрозам терроризма. – Таганрог : ЮФУ, 2009. – С. 138–146.

34 **Шарий, Т. В.** О проблеме параметризации речевого сигнала в современных системах распознавания речи / Т. В. Шарий // Вісник Донецького національного університету. – Сер. А: Природничі науки. – Вип. 2. – 2008. – С. 536–541.

35 **Старченко, И. Б.** Практикум по курсу «Математическое моделирование биологических процессов и систем» / И. Б. Старченко, В. Ю. Вишневецкий. – Таганрог : ТТИ ЮФУ, 2010. – 36 с.

36 **Сэломон, Д.** Сжатие данных, изображений и звука / Д. Сэломон. – М. : Техносфера, 2004. – 368 с.

37 **Rodman, J.** The effect of bandwidth on speech intelligibility / J. Rodman. – Pleasanton, CA, USA : Polycom, 2006. – 9 p.

38 **Humes, L. E.** Understanding the speech-understanding problems of the hearing impaired / L. E. Humes // Journal of the American Academy of Audiology. – 1991. – Vol. 2. – No. 2. – P. 59–69.

39 **Hansen, C. H.** Fundamentals of acoustics / C. H. Hansen // Occupational Exposure to Noise: Evaluation, Prevention and Control. World Health Organization Special Report S64 / Dortmund, Germany : Federal Institute for Occupational Safety and Health, 2001. – P. 23–52.

40 **Chu, W. C.** Speech coding algorithms: Foundation and evolution of standardized coders / W. C. Chu. – Hoboken, New Jersey, USA : John Wiley & Sons, 2003. – 584 p.

41 **Бочаров, И. В.** Распознавание речевых сигналов на основе метода спектрального оценивания [Электронный ресурс] / И. В. Бочаров, Д. Ю. Акатьев // Исследовано в России. – 2003. – № 6. – С. 1537–1546. – Режим доступа: http://zhurnal.ape.relarn.ru/articles/2003/130.pdf

42 **Campbell, J. P. Jr** Speaker recognition: a tutorial / J. P. Campbell Jr // Proceedings of the IEEE. – 1997. – Vol. 85. – No. 9. – P. 1437–1462.

43 **Марпл, С. Л.** Цифровой спектральный анализ и его приложения / С. Л. Марпл. – М. : Мир, 1990. – 265 с.

44 **Рогалев, А. Ф.** Основы лингвистических знаний : учеб. пособие / А. Ф. Рогалев. – Гомель : УО «Гомельский государственный университет имени Франциска Скорины», 2013. – 221 с.

45 **Кузьмина, О. Д.** Языкознание : учебное пособие / О. Д. Кузьмина, О. Ю. Макарова, О. В. Акимова, А. Е. Астафьева. – Казань : КГМУ, 2012. – 54 с.

46 **Ткач Т. Г.** Описание состава гласных фонем в аспекте обучения русскому языку как иностранному / Т. Г. Ткач // Педагогическое образование в России. – 2011. – № 1. – С. 170–175.

47 **Мильруд, Р. П.** Символизация культуры в языке / Р. П. Мильруд // Научный диалог-2012 / Екатеринбург. – 2012. – № 10. – С. 127–151.

48 **Ronzhin, A. L.** Large vocabulary automatic speech recognition for Russian language / A. L. Ronzhin, A. A. Karpov // Proc. of Second Baltic Conference on Human Language Technologies / Tallinn, Estonia. – 2005. – P. 329–334.

49 **Косарев, Ю. А.** Естественная форма диалога с ЭВМ / Ю. А. Косарев. – Л. : Машиностроение. Ленингр. отд-ние, 1989. – 143 с.

50 **Николенко, Л. А.** Формирование признаков для дикторонезависимого распознавания фонем русского языка / Л. А. Николенко // Материалы Всероссийской научно-методической конференции «Повышение качества высшего профессионального образования» / Красноярск : ИПК СФУ, 2008. – Ч. 2. – С. 323–326.

51 **Князев, С. В.** Современный русский литературный язык: Фонетика, орфоэпия, графика и орфография : учебное пособие / С. В. Князев, С. К. Пожарицкая. – 2-е изд., перераб. и доп. – М. : Академический Проект; Гаудеамус, 2011. – 430 с.

52 **Волокитин, А. А.** Параметрическое описание речевого сигнала / А. А. Волокитин, В. П. Бондаренко // Сборник докладов Научной сессии ТУСУР-2005 / Томск : ТУСУР, 2005. – С. 211–213.

53 **Петров, А. А.** Выделение признаков речевого сигнала на основе вейвлет-анализа / А. А. Петров // Сборник трудов VI Всероссийской научно-практической конференции Молодежь и современные информационные технологии / Томск : ТПУ, 2008. – С. 135–136.

54 **Продеус, А. Н.** Частотное распределение формант украинской и русской речи / А. Н. Продеус // Электроника и связь. – 2009. – № 6. – С. 18–25.

55 **Klatt, D. H.** Linguistic uses of segmental durations in English: acoustic and perceptual evidence / D. H. Klatt // Journal of the Acoustical Society of America (JASA). – 1976. – № 7. – P. 1208.

56 **Klatt, D. H.** Synthesis by rule of segmental durations in English sentences / D. H. Klatt. – N.Y. : Academic Press, 1979. – 287 p.

57 **Cooper, W. E.** Syntactic control of speech timing / W. E. Cooper. Ph. D. Thesis. – MIT, 1975.

58 **Lindblom, B.** Duration patterns of Swedish phonology: do they reflect shortterm motor memory process? / B. Lindblom, B. Lyberg, K. Holmgren. – Stockholm : Rep. Stockholm Univ., 1977. – 17 p.

59 **Klatt, D. H.** A strategy for the perceptual interpretation of duration cues in English sentences / D. H. Klatt // Working Papers. – MIT, SCG. – 1982. – Vol. 1. – P. 83.

60 **Цыплихин, А. И.** Двумерные распределения фонетических сегментов / А. И. Цыплихин, А. С. Леонов, В. Н. Сорокин // Труды Международного семинара «Диалог» / Протвино. – 2002. – С. 484–495.

61 **Алдошина, И. А.** Основы психоакустики. Часть 17. Слух и речь. Часть 1 / И. А. Алдошина // Звукорежиссер / М., 2002. – № 1. – С. 38–43.

62 **Johnson, K.** Acoustic and Auditory Phonetics / K. Johnson. – 3rd Edition. – Malden, MA : Wiley-Blackwell, 2012. – 232 p.

63 **Галунов, В. И.** Акустическая теория речеобразования и системы фонетических признаков / В. И. Галунов, В. И. Гарбарук // Материалы международной конференции «100 лет экспериментальной фонетике в России» / СПб. : СПбГУ, 2001. – С. 58–62.

64 **Ali, A. A.** An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants / A. A. Ali, J. van der Spiegel, P. Mueller // Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing. – 1998. – P. 961–964.

65 **Аграновский, А. В.** Система автоматической классификации фонем русского языка при ее обучении методом группового учета аргументов /

А. В. Аграновский, Д. А. Леднов, С. А. Репалов, Б. А. Телеснин // Искусственный интеллект. – 2000. – № 3. – С. 400–403.

66 **Жуйков, В. Я.** Алгоритм автоматической классификации сегментов речи на основе автокорреляционных и энергетических характеристик / В. Я. Жуйков, Н. Н. Кузнецов, А. Н. Харченко // Электроника и связь. – Тематический выпуск «Электроника и нанотехнологии». – 2010. – № 5. – С. 83–89.

67 **Дорохин, О. А.** Сегментация речевого сигнала / О. А. Дорохин, Д. Г. Старушко, Е. Е. Федоров, В. Ю. Шелепов // Искусственный интеллект. – 2000. – № 3. – С. 450–458.

68 **Крашенинникова, Н. А.** Основные факторы, мешающие распознаванию речевых команд / Н. А. Крашенинникова // Симбирский научный вестник. – 2011. – № 1 (3). – С. 188–191.

69 **Huang, X.** Language Processing: A guide to theory, algorithm, and system development / X. Huang, A. Acero, H. Hon. – Prentice Hall, 2001. – 1008 p.

70 **Первушин, Е. А.** Система идентификации дикторов на основе объединения признаков, векторного квантования и нормализации расстояний / Е. А. Первушин // Фундаментальные исследования. – 2011. – № 12. – Ч. 1. – С. 151–154.

71 **Hermansky, H.** Perceptual Linear Predictive (PLP) Analysis of Speech / H. Hermansky // The Journal of the Acoustical Society of America. – 1990. – Vol. 87 (4). – P. 1738–1752.

72 **Sen, S.** Design of Intelligent Control System Using Acoustic Parameters for Grinding Mill Operation / S. Sen, A. Bhaumik // National Conference on Advancement of Computing in Engineering Research (ACER-13) / Krishnagar, West Bengal, India. – 2013. – P. 261–268.

73 **Ахмад Х. М.** Математические модели принятия решений в задачах распознавания говорящего / Х. М. Ахмад // Вестник ТГТУ. – 2008. – Т. 14. – № 1. – С. 19–32.

74 **Barabanov, A. E.** Allophone segmentation by cepstra statistics / A. E. Barabanov, P. V. Moskalevich // Proc. of the Ninth International Conference «Computer Data

Analysis and Modeling» / Minsk : Belarusian State University, 2010. – Vol. 2. – P. 186–189.

75 **Tan, B. T.** The use of wavelet transforms in phoneme recognition / B. T. Tan, M. Fu, A. Spray, P. Dermody // Proc. of International Conference on Spoken Language Processing (ICSLP). – 1996. – Vol. 4. – P. 2431–2434.

76 **Dusan, S.** On the Relation Between Maximum Spectral Transition Positions and Phone Boundaries / S. Dusan, L. R. Rabiner // Proc. of ICSLP. – 2006. – P. 17–21.

77 **Шарий, Т. В.** Об одном методе автоматической сегментации речевых сигналов / Т. В. Шарий // Бионика интеллекта: научно-технический журнал. – 2009. – № 2 (71). – С. 61–65.

78 **Basile, P.** The time-scale transform method as an instrument for phonetic analysis / P. Basile, F. Cutugno, P. Maturi, A. Piccialli // Visual representations of speech signals / Chicester, UK : John Wiley & Sons, 1993. – Chapter 13. – P. 169–174.

79 **Yermolenko, T. V.** Segmentation of a speech signal with application of fast wavelet-transformation / T. V. Yermolenko // International Journal on Information Theories and Applications. – 2003. – Vol. 10. – No. 3. – P. 306–310.

80 **Ziolko, B.** Wavelet method of speech segmentation / B. Ziolko, S. Manandhar, R. Wilson, M. Ziolko // Proceedings of 14th European Signal Processing Conference EUSIPCO / Florence, Italy. – 2006.

81 **Kronland-Martinet, R.** Analysis of sound patterns through wavelet transforms / R. Kronland-Martinet, J. Morlet, A. Grossmann // International Journal of Pattern Recognition and Artificial Intelligence. – 1987. – No. 1 (2). – P. 273–302.

82 **Gerhard, D.** Pitch extraction and fundamental frequency: history and current techniques / D. Gerhard. – Regina, Saskatchewan, Canada : University of Regina. – 2003. – 22 p.

83 **Кодзасов, С. В.** Фонетическая база данных ИРЯ РАН как источник просодических сведений / С. В. Кодзасов // Просодический строй русской речи / М. : Институт русского языка РАН, 1996. – 256 с.

84 **Talkin, D.** A robust algorithm for pitch tracking (RAPT) / D. Talkin // Speech Coding and Synthesis (W. B. Kleijn and K. K. Paliwal, eds.). – ch. 14. – Elsevier Science. – 1995. – P. 495–518.

85 **Азаров, И. С.** Алгоритм оценки мгновенной частоты основного тона речевого сигнала / Азаров И. С., Вашкевич М. И., Петровский А. А. // Цифровая обработка сигналов. – 2012. – № 4. – С. 49–57.

86 **Бочаров, И. В.** Распознавание речевых сигналов на основе корреляционного метода [Электронный ресурс] / И. В. Бочаров, Д. Ю. Акатьев // Исследовано в России. – 2003. – № 6. – С. 1547–1557. – Режим доступа: http://zhurnal.ape.relarn.ru/articles/2003/131.pdf

87 **Огородников, А. Н.** Эффективный алгоритм оценивания длины периода основного тона речевого сигнала / А. Н. Огородников // Материалы VIII Всеросс. научн.-практ. конф. «Научное творчество молодежи» / Томск : Изд-во Тос. ун-та, 2004. – С. 52–53.

88 **Попов, В. И.** Основы сотовой связи стандарта GSM / В. И. Попов. – М. : Эко-Трендз, 2005. – 296 с.

89 **Рысин, Ю. С.** Влияние пауз при передаче сложносоставных числительных по IP-сетям связи на коэффициент эффективных попыток вызова / Ю. С. Рысин, А. Н. Терехов // Материалы Международной научно-технической конференции INTERMATIC / М. : МИРЭА, 2012. – Ч. 5. – С. 98–103.

90 G.729, Annex B, A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70 / ITU-T Recommendation, 1996.

91 **Benyassine, A.** ITU-T recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications / A. Benyassine, E. Shlomot, H-Y Su // IEEE Communications Magazine. – 1997. – No. 35 (9). – P. 64–73.

92 **Farsi, H.** Improving voice activity detection used in ITU-T G.729.B / H. Farsi, M. A. Mozaffarian, H. Rahmani // Proceedings of the 3rd WSEAS International Conference on Circuits, Systems, Signal and Telecommunications (CISST'09). – 2009. – P. 11–15.

93 **Villavicencio, F.** Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation / F. Villavicencio, A. Röbel and X. Rodet // Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) / Toulouse , France. – 2006. – Vol. 1. – P. 869–872.

94 **Vijayachandran, V. M.** A novel algorithm for voice activity detection / V. M. Vijayachandran, K. B. Shobha Devi // Proc. of WSES/IEEE International Multiconference: Speech, Signal and Image Processing / Malta. – 2001.

95 **Sang-Sik, A.** An improved statistical model-based VAD algorithm with an adaptive threshold / A. Sang-Sik, L. Yoon-Chang // Journal of The Chinese Institute of Engineers / Taipei. – 2006. – Vol. 29. – No. 5. – P. 783–789.

96 **Mermelstein, P.** Automatic segmentation of speech into syllabic units / P. Mermelstein // Journal of the Acoustical Society of America. – 1975. – Vol. 58. – N. 4. – pp. 880-883.

97 **Огородников, А. Н.** Выбор интервалов анализа сигнала при распознавании речи / А. Н. Огородников, В. В. Поддубный // Тез. докл. Десятой Международной научно-технической конференции студентов и аспирантов Радиоэлектроника, электротехника и энергетика / М. : МЭИ, 2004. – Т. 1. – С. 291–292.

98 **Mas Soro, P.** A spectral estimator of vocal jitter / P. Mas Soro, J. Schoentgen // Brussel, Belguim : Université libre de Bruxelles, 2011. – 108 p.

99 **Михайлов, В. Г.** Из истории исследований преобразования речи / В. Г. Михайлов // Речевые технологии. – 2008. – № 1. – С. 93–113.

100 **Кодзасов, С. В.** Общая фонетика / С. В. Кодзасов, О. Ф. Кривнова. – М. : Рос. гос. гуманит. ун-т, 2001. – 592 с.

101 **Женило, В. Р.** Исследование вибрато голоса / В. Р. Женило // Труды Международной конференции «Информатизация правоохранительных систем» / М., 1999. – С. 335–337.

102 **Архипов, И. О.** Оценка точности выделения основного тона методом GS / И. О. Архипов, В. Б. Гитлин // Современные речевые технологии. Сборник

трудов IX сессии Российского акустического общества / М. : ГЕОС, 1999. – С. 38–42.

103 **Ахмад, Х. М.** Определение высоты тона методом произведения гармоник спектра речевого сигнала / Х. М. Ахмад // Вестник ТГТУ / Тамбов : Изд-во ТГТУ, 2007. – Т. 13. – № 3. – С. 712–714.

104 Вокодерная телефония. Методы и проблемы / под ред. А. А. Пирогова. – М. : Связь, 1974. – 536 с.

105 **Андрейченко, Л. Н.** Русский язык. Фонетика и фонология. Орфоэпия. Графика и орфография / Л. Н. Андрейченко; под ред. Г. Г. Инфантовой и Н. А. Сениной. – М. : Флинта, 2003. – 231 с.

106 Энциклопедический словарь / Репр. воспр. изд. Ф. А. Брокгауз – И. А. Ефрон 1890 г. – М. : Терра, 2001. – 40726 с.

107 **Фант, Г.** Акустическая теория речеобразования / Г. Фант; под ред. В. С. Григорьева. – М. : Наука, 1964. – 284 с.

108 **Галяшина, Е. И.** Речь под микроскопом / Е. И. Галяшина // Компьютерра. – 1999. – № 15 (293). – С. 16–24.

109 **Свириденко, В. А.** Аутентификация личности по голосу [Электронный ресурс] // Мобильные системы. – 2004. – № 2. – Режим доступа: http://web.archive.org/web/20071202155053/http://www.spirit.ru/articles/svi_mb.html

110 **Sukittanon, S.** Modulation-scale analysis for content identification / S. Sukittanon, L. E. Atlas, J. W. Pitton // IEEE Transactions On Signal Processing. – 2004. – Vol. 52. – No. 10. – P. 3023–3035.

111 **Зилинберг, А. Ю.** Разработка и исследование временных и спектральных алгоритмов VAD (Voice Activity Detection) / А. Ю. Зилинберг, Ю. А. Корнеев // Российская школа-конференция «Мобильные системы передачи данных» / Зеленоград : МИЭТ, 2006. – С. 58–70.

112 **Beritelli, F.** Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors / F. Beritelli, S. Casale, G. Ruggeri, S. Serrano // IEEE Signal Processing Letters. – 2002. – Vol. 9 (3). – P. 85–88.

113 **Stejskal, V.** Empty speech pause detection algorithms' comparison / V. Stejskal, N. Bourbakis, A. Esposito // International Journal of Advanced Intelligence. – 2010. – Vol. 2. – No. 1. – P. 145–160.

114 **Majeed, S. A.** Mel frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: a comparison study / S. A. Majeed, H. Husain, S. A. Samad, T. F. Idbeaa // Journal of Theoretical and Applied Information Technology. – 2015. – Vol. 79. – No. 1. – P. 38–56.

115 **Xugang, L.**, Lateral inhibition mechanism in computational auditory model and its application in robust speech recognition / L. Xugang, L. Gang, W. Lipo // Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop. – 2000. – Vol. 2. – P. 785–794.

116 ISO/IEC International Standard IS 11172-3 "Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbits/s - Part 3: Audio". – 1993.

117 **Premananda, B. S.** Incorporating Auditory Masking Properties for Speech Enhancement in presence of Near-end Noise / B. S. Premananda, B. V. Uma // International Journal of Computer Applications, IJCA. – 2014. – Vol. 106. – No. 15. – P. 1–6.

118 **Painter, T.** Perceptual Coding of Digital Audio / T. Painter, A. Spanias // Proceedings of the IEEE. – 2000. – Vol. 88. – No. 4. – P. 451–513.

119 **Dai, P.** An improved model of masking effects for robust speech recognition system / P. Dai, Y. Soon // Speech Communication. – 2013. – Vol. 55. – P. 387–396.

120 **Lee, L. M.** HMM Speech Recognition in Matlab [Электронный ресурс] / L. M. Lee // 2015. – Режим доступа: http://sourceforge.net/projects/hmm-asr-matlab/

121 **Lee, L. M.** Duration High-Order Hidden Markov Models and Training Algorithms for Speech Recognition / L. M. Lee // Journal of Information Science and Engineering. – 2015. – Vol. 31. – No. 3. – P. 799–820.

122 **Gonzalez, S.** PEFAC - a pitch estimation algorithm robust to high levels of noise / S. Gonzalez, M. Brookes // IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). – 2014. – Vol. 22. – No. 2. – P. 518–530.

123 **Dai, P.** A temporal frequency warped (TFW) 2D psychoacoustic filter for robust speech recognition system / P. Dai, Y. Soon // Speech Communication. – 2011. – Vol. 53. – P. 229–241.

124 **Noll, P.** MPEG Digital Audio Coding Standards / P. Noll; The Digital Signal Processing Handbook. Edited by V. K. Madisetti and D. B. Williams. – IEEE Press/CRC Press, 1998. – P. 40-1 – 40-28.

125 **Зилинберг, А. Ю.** Разработка и исследование алгоритмов многоуровневой временной сегментации речевых сигналов: диссертация … кандидата технических наук. – СПб., 2010. – 161 с.

126 **Томчук, К. К.** Разработка и исследование алгоритма модификации темпа произнесения речи: диссертация … магистра техники и технологии. – СПб., 2009. – 109 с.

127 **Sohn, J.** A statistical model-based voice activity detection / J. Sohn, N. S. Kim, W. Sung // IEEE Signal Processing Letters. – 1999. – Vol. 6. – No. 1. – P. 1–3.

128 **Хованова, Н. А.** Методы анализа временных рядов / Н. А. Хованова, И. А. Хованов. – Саратов : ГосУНЦ «Колледж», 2001. – 120 с.

129 **Drugman, T.** Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review / T. Drugman, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit // IEEE Transactions on Audio, Speech, and Language Processing. – 2012. – Vol. 20. – No. 3. – P. 994–1006.

130 **Sujith, P.** An Error Correction Scheme for GCI Detection Algorithms using Pitch Smoothness Criterion / P. Sujith, A. Prathosh, A. Ramakrishnan, P. Ghosh // Proc. 16th Intern. Conf. INTERSPEECH, Dresden, Germany. – 2015. – P. 3284-3288.

131 **Kane, J.** Evaluation of glottal closure instant detection in a range of voice qualities / J. Kane, C. Gobl // Speech Communication. – 2013. – Vol. 55. – No. 2. – P. 295-314.

132 **Drugman, T.** Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics / T. Drugman, A. Alwan // Proc. 11th Intern. Conf. INTERSPEECH, Firenze, Italy. – 2011. – P. 1973–1976.

133 **Калинцев, Ю. К.** Разборчивость речи в цифровых вокодерах / Ю. К. Калинцев. – М. : Радио и связь, 1991. – 220 с.

134 **Бабкин, А. В.** Оценка качества системы синтеза речи, разработанного в МГУ. / А. В. Бабкин, Л. М. Захаров. // Труды международного семинара Диалог’99 по компьютерной лингвистике и ее приложениям. – Таруса, 1999. – С. 12–25.

135 **Бабкин, А. В.** Автоматический синтез речи – проблемы и методы генерации речевого сигнала / А. В. Бабкин // Труды международного семинара Диалог’98 по компьютерной лингвистике и ее приложениям. – Казань, 1998. – С. 425–437.

136 **Прохоров, Ю. К.** Управленческие решения : учебное пособие / Ю. К. Прохоров, В. В. Фролов. – 2-е изд., испр. и доп. – СПб. : СПбГУ ИТМО, 2011. – 138 с.

137 **Монахова, М. М.** Модели и алгоритмы контроля инцидентов информационной безопасности в корпоративной телекоммуникационной сети: диссертация … кандидата технических наук. – Владимир., 2016. – 137 с.

ПРИЛОЖЕНИЕ А

Методика исследования сигнальных особенностей звуков

А.1 Программное обеспечение исследования

При организации исследования сигнальных особенностей звуков русской речи необходимо учесть необходимость ряда немаловажных факторов и условий:

- достаточное количество исследуемых аллофонов;
- достаточное количество исследуемых фонограмм;
- использование фонограмм, записанных по разным дикторам;
- рассмотрение различных групп звуков;
- рассмотрение абсолютного положения звука в слове, а также относительного положения в окружении других типов звуков;
- получение текущей статистики на разных этапах исследования;
- возможность внесения в промежуточные результаты корректировок вплоть до уровня одной реализации звука.

Перечисленные выше положения, в итоге, приводят к сложной структуре и глубокой взаимосвязи всего разнообразия промежуточных данных исследования. В связи с этим наиболее подходящим решением является использование специализированной базы данных. В рамках исследования для этой цели используется система управления базами данных MySQL.

В исследовательской компьютерной среде MATLAB, используемой для технической обработки исследуемых данных, моделирования, вывода графической информации, также имеются встроенные средства интеграции с базой данных MySQL. Однако, ввиду сравнительной сложности организации интеграции, а также малого акцентирования на вопросах работы с файловой системой и обработки строковых данных, в особенности кириллических символов, в качестве связующего MySQL и MATLAB звена в исследовании задействован язык серверных приложений PHP. Язык PHP (англ. PHP: Hypertext Preprocessor) изначально хорошо сочетается с распространенным серверным программным обеспечением, необходимым для его работы – Apache. В PHP на

высоком уровне организована работа с базой данных MySQL, а также имеется исчерпывающий набор встроенных функций для обработки строк и многомерных массивов различных типов данных.

Взаимодействие MATLAB и PHP осуществляется с помощью широко распространенного протокола передачи данных HTTP. Общая структура обмена данными между перечисленными программными средствами представлена в виде структурной схемы на рисунке А.1.

Рисунок А.1 – Структура взаимодействия программных средств при работе с данными

А.2 Подготовка базы слов для исследования

Для исследования сигнальных особенностей звуков русской речи необходимо подобрать перечень слов, в которых в итоге встречались бы все выявленные на основе русского фонетического алфавита звуки. При этом, чем большее количество реализаций каждого аллофона будет присутствовать в базе,

тем, очевидно, более состоятельные статистические характеристики по каждому звуку будут получены. В исследовании была поставлена цель обеспечить появление каждого аллофона из базового перечня у каждого диктора и в определенной позиции не менее двух раз. К примеру, звук [п’] присутствует в виде 14 реализаций: для двух дикторов по 3 в начале слова, по 2 в середине слова и по 2 в конце слова. Наиболее часто встречающиеся в речи аллофоны, например, [а], имеют подобным образом в базе до 114 реализаций.

Для обеспечения приемлемой оценки «дикторозависимости» параметров все слова произносились двумя дикторами: мужского и женского пола.

В целях поиска слов с наиболее редко встречающимися звуками, а также для автоматизации определения гласного звука в сильной позиции (ударного гласного) был использован словарь ударений русского языка, содержащий информацию о 163290 словах.

Для последующей ручной сегментации была подготовлена фонетическая транскрипция для всех подобранных для исследования слов. Данный процесс частично автоматизирован за счет программной реализации базовых правил произношения, то есть трансформации буквенного написания слов в последовательность звуков.

В частности, учтены следующие базовые правила произношения:

- аллофон [ы] в буквосочетаниях *ЖИ*, *ШИ*, *ЦИ;*
- смягчение согласного в буквосочетаниях *ВИ*, *ДИ*, *КИ*, *ЛИ*, *МИ*, *НИ;*
- смягчение предыдущей согласной гласными *Е*, *Ё*, *Ю*, *Я* (кроме *Ц*, у которой нет мягкого варианта произнесения, и *Ж*);
- преобразование *Е* после мягкого согласного под ударением в аллофон [э];
- оглушение согласных на конце слов (*Б* → [п], *В* → [ф], *Г* → [к] и т.д.);
- буквосочетания *ЗЧ*, *ЖЧ*, *СЧ*, *СЩ* обычно произносятся как [щ'];
- буквосочетания *ДТ*, *ТЧ* обычно произносятся как [ч'];
- звонкие согласные перед глухими оглушаются;
- *Ь* смягчает предыдущую согласную;

– буква *Й* представляется аллофоном [j];
– буквы *Ю, Я, Ё*, следующие после гласной или *Ь*, представляются двумя аллофонами, соответственно: [jу], [ja] и [jo];
– буквы *Ю, Я, Ё*, следующие после согласной, смягчают ее и представляются аллофонами, соответственно: [у] , [а] и [о];
– буквы *А* и *О* во второй слабой позиции в начале слова представляются аллофоном [^];
– гласные *А, О, Э, Е, Я* во второй слабой позиции после мягких согласных представляются аллофоном [ь], а после твердых – [ъ];
– гласные *А, О, Э* в первой слабой позиции после твердых согласных представляются аллофоном [^];
– буквы *А, Е* в предударном после твердых согласных *Ж, Ш, Ц* произносятся как аллофон [ы^{э}];
– буквы *А, О, Э, Е* в предударном после мягких согласных произносятся как аллофон [и^{э}];
– буква *Е* под ударением после твердых согласных *Ж, Ш, Ц* или после мягких согласных представляется аллофоном [э];
– звуки [щ’] и [ч’] всегда мягкие, поэтому буквы *Щ* и *Ч* всегда смягчаются.

Реализация перечисленных закономерностей позволяет в большой степени автоматизировать процесс транскрибирования. Примеры результатов работы алгоритма транскрибирования русских слов в последовательность звуков для 10 взятых подряд из базы исследования слов показаны в таблице А.1

Таблица А.1. Автоматизация транскрибирования слов

Слово	Автоматическая транскрипция	Действительная транскрипция	Число правок от исходного слова	Число правок от автоматической транскрипции
бел	[б'|Э|л]	[б'|Э|л]	2	0
бела	[б'|иэ|л|А]	[б'|иэ|л|А]	3	0
большого	[б|^|л'|ш|О|г|ъ]	[б|^|л'|ш|О|в|ъ]	5	1

бреют	[б'|р'|Э|j|у|т]	[б|р'|Э|j|у|т]	4	1
вафли	[в|А|ф'|л'|и]	[в|А|ф|л'|и]	2	1
весть	[в'|Э|с'|т']	[в'|Э|с'|т']	4	0
взвод	[в|з|в|О|т]	[в|з|в|О|т]	2	0
взял	[в'|з'|А|л]	[в|з'|А|л]	2	1
взяла	[в'|з'|а|л|А]	[в|з'|иэ|л|А]	3	2
воз	[в|О|с]	[в|О|с]	1	0
Всего 10 слов			28	6

Таким образом, для исследования была подготовлена база из 184 слов и их транскрипций. В таблице А.2 для этой базы приведена верхняя часть (только основные гласные звуки) таблицы с количеством вхождений определенных звуков в определенных позициях. Позиции, в которых звук не может встречаться, обозначены прочерками: например, звук [а] в слабой позиции не существует. В разработанном программном интерфейсе отображения данной таблицы реализован также вывод соответствующего списка слов с транскрипцией при наведении на определенную позицию (рисунок А.2).

Таблица А.2. Количество вхождений основных гласных в базу исследования

Звук	Безударный			Ударный		
	В начале слова	В середине слова	В конце слова	В начале слова	В середине слова	В конце слова
[и]	1	12	2	3	16	2
[ы]	-	4	5	-	7	2
[а]	-	-	-	2	47	8
[о]	-	-	-	3	40	2
[э]	-	-	-	4	31	2
[у]	6	7	2	2	9	4

Звук	Безударный			Ударный		
	В начале слова	В середине слова	В конце слова	В начале слова	В середине слова	В конце слова
п'	3	2	2	-	-	-
р	3	26	7	-	-	-
р'	2	14	3			-
с	9	19	9			-
с'	6	7	3			-
т	3	20	8	-	-	-

автор [А|ф|т|ъ|р]
кефир [к'|иэ|ф'|И|р]
пар [п|А|р]
парикмахер [п|ъ|р'|ие|к|м|А|х'|ь|р]
сэр [с|Э|р]
фетр [ф'|Э|т|р]
шар [ш|А|р]

Рисунок А.2 – Отображение количества вхождений звуков в слова для исследования в программном интерфейсе

А.3 Обработка записанных речевых фонограмм

Для обработки исходная фонограмма должна быть представлена в цифровом виде, то есть РС должен быть дискретизирован по времени и квантован по амплитуде. В этом же блоке предварительной обработки фонограмм могут производиться и другие операции, например, нормализация и устранение смещения по постоянному току. Нормализация приводит средний уровень громкости разных фонограмм к одному значению. В частности, при хранении фонограмм в распространенном формате WAV (сокращение от Waveform, «форма сигнала») величина сигнала может менять свои значения в диапазоне от *-1* до *+1*. Самый простой способ нормализации заключается в поиске пика максимального уровня в фонограмме и усиления всей фонограммы на величину этого пика, так чтобы пик принял значение *0* дБ. При этом дальнейшее увеличение уровня фонограммы приведет к ее клиппированию (clipping, ограничение амплитуды) – перегрузке, влекущей нежелательные искажения, которые также хорошо заметны на слух.

Звуковое оборудование может вносить в РС сдвиг по постоянному току[1]. Наличие смещения по постоянному току создает две проблемы обработки РС. Во-

[1] Загуменнов, А. П. Запись и редактирование звука. Музыкальные эффекты. – М. : Издательство «НТ Пресс», 2005. – 181 с.

первых, при конкатенации фрагментов из разных записей нарушается гладкость соединения. Во-вторых, некоторые функции обработки звука некорректно срабатывают при наличии в материале смещения по постоянному току. Для коррекции смещения из каждого отсчета PC вычитается среднее арифметическое значение всех отсчетов сигнала[2]:

$$\bar{x}_j = x_j - \frac{1}{N}\sum_{i=0}^{N-1} x_{i,} \quad j = \overline{0, N-1}, \tag{A.1}$$

где N – количество отсчетов сигнала, x_j – j-ый отсчет исходного сигнала со смещением, $\bar{x}_j$ – j-ый отсчет корректированного сигнала.

[2] Котомин, А. В. Предобработка звукового сигнала в системе распознавания речевых команд. // Труды XV Молодежной научно-практической конференции Наукоёмкие информационные технологии: SIT-2011 / Переславль-Залесский : Изд-во «Университет города Переславля», 2011. – С. 25–38.

ПРИЛОЖЕНИЕ Б

Дополнительные таблицы и диаграммы к результатам исследования сигнальных особенностей звуков русской речи

Таблица Б.1. Средние по дикторам длительности звуков, без учета ударности гласного, положения звука в слове, длительности остаточных колебаний связок

Звук	Средняя длительность без учета остаточных осцилляций, мс	СКО, мс	Кол-во измерений	Мин., мс	Макс., мс
[ж:]	229,6	50,6	6	157,1	291,9
[щ']	228,5	44,9	21	144,0	320,9
[д:]	222,6	19,3	2	208,9	236,2
[а]	216,0	48,7	112	113,0	332,2
[о]	212,1	51,2	90	107,8	329,3
[э]	202,7	53,2	75	68,0	323,3
[ж':]	180,3	79,6	3	91,2	244,6
[с']	175,3	47,8	32	80,5	241,4
[ш]	175,1	39,5	23	94,2	237,7
[ы]	173,7	54,0	36	64,7	295,7
[н:]	171,1	68,2	2	122,9	219,4
[б']	170,6	50,0	6	116,9	260,3
[ж]	163,8	32,0	22	107,4	234,8
[х']	162,7	26,7	14	99,2	203,8
[у]	157,7	60,4	59	53,9	280,5
[и]	148,2	47,3	75	53,6	263,8
[с]	147,5	51,2	74	33,5	262,0
[ф']	141,2	33,4	14	78,4	188,4
[ц]	138,6	37,3	22	67,5	203,8
[з']	132,6	49,9	16	76,6	230,6
[м']	129,2	40,2	20	43,1	195,8
[х]	126,2	36,4	24	87,4	235,2
[з]	120,2	29,7	20	76,2	176,1
[д']	119,8	45,2	22	52,3	251,7
[ч']	118,9	37,1	22	67,7	206,9
[г']	114,9	41,1	8	49,7	174,4
[ъ]	114,9	49,5	102	44,9	224,3
[г]	113,3	33,9	26	57,9	171,0

Звук	Средняя длительность без учета остаточных осцилляций, мс	СКО, мс	Кол-во измерений	Мин., мс	Макс., мс
[т']	113,0	39,0	56	38,0	186,5
[н']	112,9	40,7	32	31,9	192,9
[j]	108,6	44,9	57	39,4	269,1
[н]	105,7	30,5	45	56,6	166,9
[^]	105,6	26,6	74	48,4	169,2
[м]	100,1	28,8	46	51,1	207,0
[ь]	99,6	39,3	35	51,2	231,0
[в']	99,3	51,4	16	42,6	211,2
[ф]	98,8	49,8	20	28,2	222,2
[л]	98,5	27,2	63	31,7	148,5
[д]	98,3	33,0	44	35,0	190,1
[в]	93,5	37,1	30	35,7	199,4
[и^{э}]	93,0	29,9	31	46,4	183,9
[л']	90,9	36,4	30	41,5	205,4
[ы^{э}]	89,4	22,1	11	65,2	131,8
[б]	89,1	16,2	8	60,1	111,5
[и^{е}]	84,0	20,7	39	39,5	121,3
[р']	79,6	37,0	36	17,1	186,8
[р]	75,3	30,8	72	20,4	147,3
[_]	65,3	25,6	152	14,0	127,2
[к']	56,9	14,0	10	39,0	79,4
[к]	50,8	24,3	68	16,1	142,9
[т]	36,1	25,9	60	9,8	121,0
[п']	29,9	23,9	14	7,7	82,4
[п]	24,5	17,0	42	9,4	106,7

Таблица Б.2. Средние мощности ударных гласных звуков

Звук	Средняя мощность, $x10^{-3}$	СКО, $x10^{-3}$	Кол-во измерений	Мин.	Макс.
[о]	42,7	24,4	90	10,1	118,7
[а]	37,4	18,4	112	10,1	93,7
[у]	36,7	38,0	30	5,0	163,4
[э]	32,1	16,4	73	6,7	77,2
[ы]	30,7	22,6	18	11,4	105,9
[и]	23,3	23,6	44	3,9	131,4

Таблица Б.3. Средние мощности безударных гласных звуков

Звук	Средняя мощность, $x10^{-3}$	СКО, $x10^{-3}$	Кол-во измерений	Мин.	Макс.
[^]	48,2	21,1	74	11,8	124,2
[и^{э}]	35,7	27,2	31	8,8	155,6
[у]	30,6	28,3	29	4,1	110,9
[ы^{э}]	29,3	13,0	11	8,9	45,4
[ь]	27,7	23,7	35	2,3	89,0
[э]	24,0	10,9	2	16,3	31,7
[и^{е}]	22,6	19,4	39	3,5	74,2
[ъ]	21,8	15,0	102	3,0	61,1
[и]	21,0	28,2	31	2,9	127,7
[ы]	14,0	11,7	18	3,5	54,2

Таблица Б.4. Средние по дикторам частоты пересечений нуля, без учета ударности гласного, положения звука в слове, длительности остаточных колебаний связок

Звук	Частота пересечений нуля, $x10^{-3}$	СКО, $x10^{-3}$	Кол-во измерений	Мин.	Макс.
[ц]	593,6	93,7	22	386,5	727,1
[с]	556,7	103,7	74	81,9	772,3
[с']	546,9	76,2	32	369,4	688,0
[т']	470,3	98,1	56	288,1	677,9
[х']	409,8	94,0	14	153,1	521,7
[к']	379,7	79,1	10	259,1	538,6
[ф']	342,0	114,3	14	94,4	534,9
[щ']	335,5	52,4	21	258,6	429,8
[ч']	335,4	57,1	22	253,6	425,2
[ш]	331,1	70,4	23	214,0	428,4
[ф]	324,3	126,9	20	120,9	542,8
[т]	200,6	64,8	60	56,1	345,1
[з]	189,8	100,8	20	62,9	400,6
[ж':]	183,9	37,9	3	141,9	215,5
[з']	177,8	122,6	16	52,3	569,5
[ж:]	173,0	35,0	6	136,5	214,6
[х]	155,5	54,1	24	91,6	356,7
[к]	135,8	36,8	68	77,8	278,7
[ж]	128,7	38,7	22	77,8	240,4
[_]	125,7	65,9	152	33,7	383,0
[п']	123,6	48,3	14	68,5	204,9
[а]	79,9	15,7	112	38,2	117,2
[д']	75,3	39,9	22	23,3	181,5
[п]	74,0	25,1	42	33,7	118,8
[э]	69,6	22,7	75	38,7	137,0
[^]	69,2	17,1	74	30,4	117,1
[р']	63,9	24,8	36	30,2	142,8
[ь]	58,8	27,4	35	21,2	166,2
[и^{е}]	54,8	20,3	39	29,9	126,0
[р]	54,7	14,1	72	24,0	97,8
[и^{э}]	54,4	18,2	31	29,3	90,4
[и]	54,2	26,0	75	19,5	147,5

Звук	Частота пересечений нуля, $x10^{-3}$	СКО, $x10^{-3}$	Кол-во измерений	Мин.	Макс.
[ыᵌ]	53,0	18,6	11	29,5	96,0
[ъ]	52,8	12,4	102	25,1	88,3
[г']	52,1	18,0	8	34,4	78,4
[j]	52,1	29,7	57	22,5	174,2
[л']	48,6	18,8	30	25,9	98,4
[о]	48,4	6,6	90	32,2	65,1
[ы]	47,2	17,0	36	23,4	99,7
[в']	43,5	24,1	16	27,7	123,3
[л]	39,0	7,8	63	23,7	58,6
[в]	37,5	10,2	30	24,3	60,0
[у]	33,4	5,4	59	24,1	53,1
[г]	32,1	8,4	26	18,6	59,2
[б']	28,7	12,3	6	14,3	44,6
[н']	26,7	6,6	32	14,2	44,7
[д]	26,0	7,9	44	14,0	61,6
[м']	25,7	7,0	20	18,4	46,0
[м]	25,6	5,0	46	15,3	39,6
[н]	23,9	5,5	45	13,3	40,5
[б]	23,0	4,5	8	16,0	31,2
[д:]	22,9	3,9	2	20,2	25,7
[н:]	16,6	8,4	2	10,6	22,5

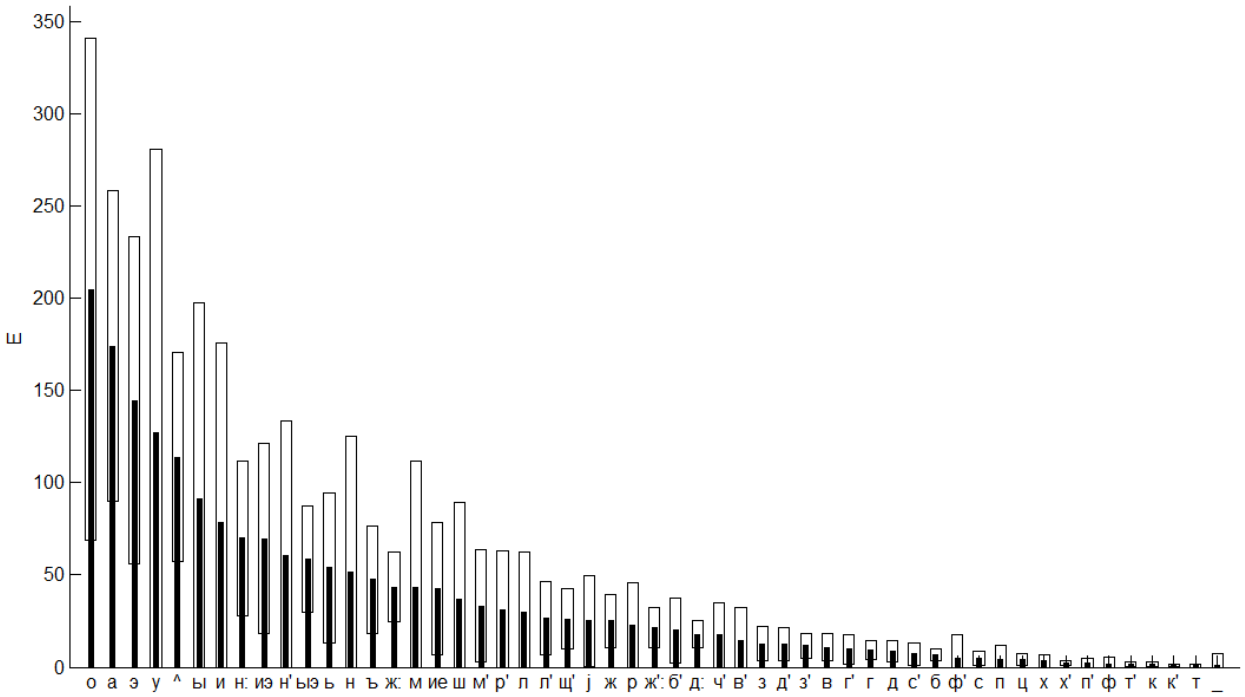


Рисунок Б.1 – Оценки средних энергий реализаций звуков ±σ-размахи

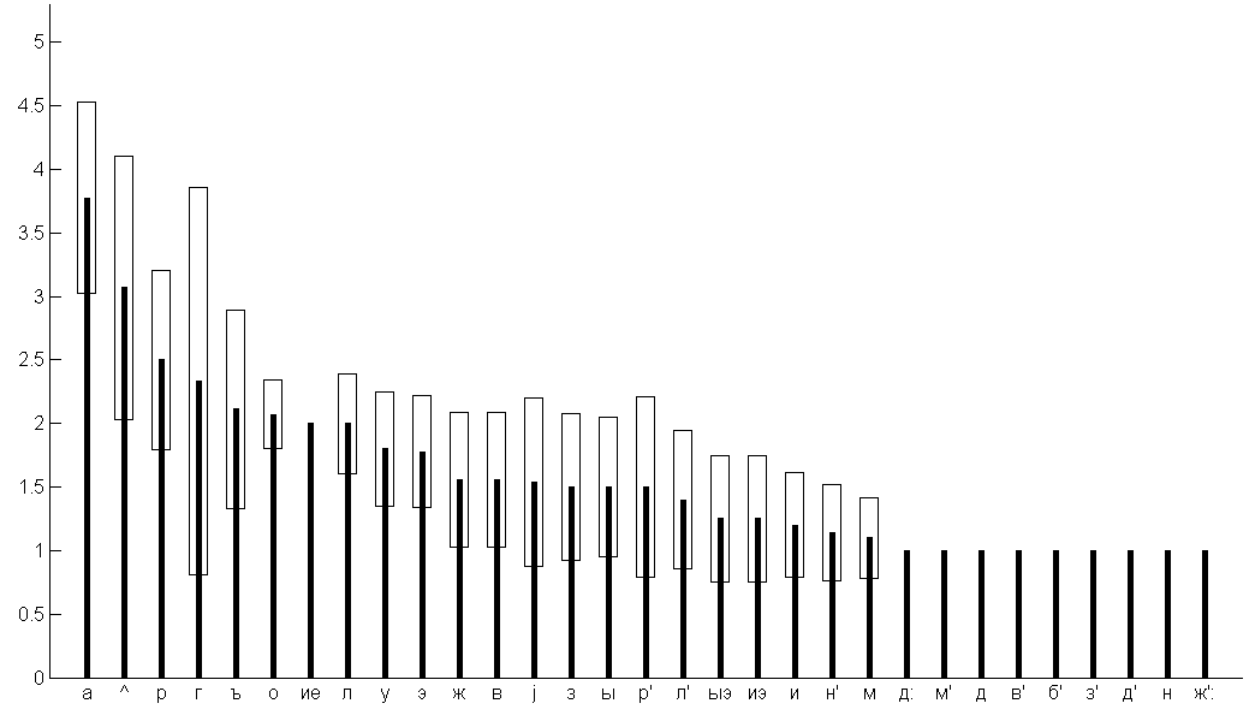


Рисунок Б.2 – Диаграмма распределения среднего количества переколебаний на периоде ОТ по звукам для исследовавшегося женского голоса ±σ-размахи

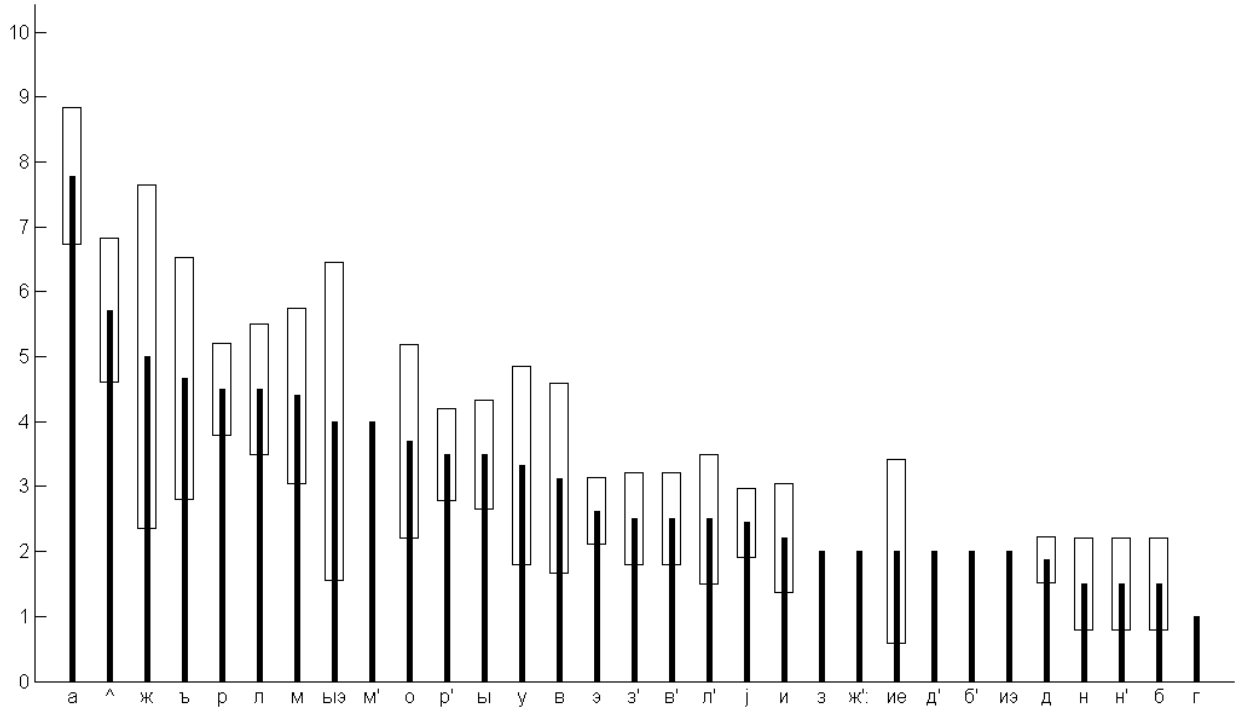


Рисунок Б.3 – Диаграмма распределения среднего количества переколебаний на периоде ОТ по звукам для исследовавшегося мужского голоса ±σ-размахи

ПРИЛОЖЕНИЕ В

Таблицы результатов распознавания одиночных слов при разных алгоритмах MFCC-параметризации

Таблица В.1. Частоты распознавания и относительные улучшения при чистом РС

алгоритм / тип шума	**MFCC(13)**	**LI**	**MPEG1**	**FFH**	**LI + FFH**	**MPEG1 + FFH**	**FFHi**	**LI + FFHi**	**MPEG1 + FFHi**
–	90,7	89,6	84,5	90,4	89,1	84,6	90,4	89,1	84,6
RI	–	-12,6	-67,8	-3,9	-17,8	-66,5	-3,9	-17,8	-66,5

Таблица В.2. Частоты распознавания и относительные улучшения при ОСШ 20 дБ

алгоритм / тип шума	**MFCC(13)**	**LI**	**MPEG1**	**FFH**	**LI + FFH**	**MPEG1 + FFH**	**FFHi**	**LI + FFHi**	**MPEG1 + FFHi**
БГШ	67,2	73,8	77,0	66,3	73,2	76,9	66,4	73,3	77,0
толпа	75,6	73,1	74,4	76,4	72,9	74,3	77,4	73,5	74,4
улица	77,6	77,0	76,6	77,7	77,4	77,0	77,9	77,5	77,2
поезд	81,1	80,5	77,9	80,9	80,7	77,5	81,0	80,6	77,3
автомобиль	68,5	69,6	74,2	68,1	69,4	74,1	68,2	69,5	74,1
среднее (ср.)	74,0	74,8	76,0	73,9	74,7	75,9	74,2	74,9	76,0
RI	–	3,0	7,8	-0,5	2,8	7,5	0,8	3,4	7,7
ср. без БГШ	75,7	75,0	75,8	75,8	75,1	75,7	76,2	75,3	75,8
RI без БГШ	–	-2,7	0,3	0,3	-2,4	0,1	1,9	-1,8	0,2

Таблица В.3. Частоты распознавания и относительные улучшения при ОСШ 15 дБ

алгоритм / тип шума	**MFCC(13)**	**LI**	**MPEG1**	**FFH**	**LI + FFH**	**MPEG1 + FFH**	**FFHi**	**LI + FFHi**	**MPEG1 + FFHi**
БГШ	49,8	60,3	67,0	48,8	60,2	67,7	49,0	60,1	67,7
толпа	66,0	62,7	69,1	67,1	62,7	69,1	69,1	63,7	69,4
улица	68,0	68,3	72,6	69,5	68,9	73,0	69,5	69,1	73,0
поезд	75,7	76,7	75,4	76,0	76,3	75,1	76,8	76,7	75,3
автомобиль	63,7	65,9	70,5	63,1	65,4	70,4	63,2	65,4	70,4
среднее (ср.)	64,7	66,8	70,9	64,9	66,7	71,1	65,5	67,0	71,2
RI	–	6,0	17,7	0,7	5,8	18,1	2,4	6,6	18,5
ср. без БГШ	68,4	68,4	71,9	68,9	68,3	71,9	69,6	68,7	72,1
RI без БГШ	–	0,1	11,2	1,8	-0,2	11,2	4,0	1,1	11,7

Таблица В.4. Частоты распознавания и относительные улучшения при ОСШ 10 дБ

алгоритм / тип шума	**MFCC(13)**	**LI**	**MPEG1**	**FFH**	**LI + FFH**	**MPEG1 + FFH**	**FFHi**	**LI + FFHi**	**MPEG1 + FFHi**
БГШ	27,6	39,5	50,8	27,6	40,1	51,6	28,0	40,4	51,6
толпа	54,0	51,2	58,8	54,2	51,3	58,5	55,8	52,4	59,3
улица	54,4	55,5	64,4	56,0	56,9	65,2	56,6	57,1	65,3
поезд	66,5	68,7	69,3	67,0	68,6	69,8	68,5	69,5	69,7
автомобиль	58,9	62,2	67,8	58,7	61,9	67,1	58,4	61,9	67,1
среднее (ср.)	52,3	55,4	62,2	52,7	55,7	62,4	53,5	56,3	62,6
RI	–	6,6	20,9	0,9	7,3	21,3	2,5	8,4	21,7
ср. без БГШ	58,4	59,4	65,1	59,0	59,7	65,1	59,8	60,2	65,4
RI без БГШ	–	2,3	16,0	1,3	2,9	16,1	3,4	4,3	16,6

Таблица В.5. Частоты распознавания и относительные улучшения при ОСШ 5 дБ

алгоритм / тип шума	**MFCC(13)**	**LI**	**MPEG1**	**FFH**	**LI + FFH**	**MPEG1 + FFH**	**FFHi**	**LI + FFHi**	**MPEG1 + FFHi**
БГШ	15,4	20,6	31,9	15,2	21,4	32,7	15,6	22,0	32,9
толпа	39,8	38,7	44,7	41,0	38,6	45,1	42,8	39,6	45,7
улица	33,5	35,6	50,8	36,0	37,9	51,7	36,6	38,1	51,8
поезд	53,1	55,1	61,5	54,4	56,2	61,9	55,8	57,4	62,0
автомобиль	51,8	56,8	63,0	52,6	56,4	62,6	52,5	56,6	62,6
среднее (ср.)	38,7	41,4	50,4	39,8	42,1	50,8	40,7	42,7	51,0
RI	–	4,3	19,0	1,8	5,5	19,7	3,2	6,5	20,0
ср. без БГШ	44,6	46,6	55,0	46,0	47,3	55,3	46,9	47,9	55,5
RI без БГШ	–	3,6	18,9	2,5	4,9	19,4	4,3	6,0	19,8

Таблица В.6. Частоты распознавания и относительные улучшения при ОСШ 0 дБ

алгоритм / тип шума	**MFCC(13)**	**LI**	**MPEG1**	**FFH**	**LI + FFH**	**MPEG1 + FFH**	**FFHi**	**LI + FFHi**	**MPEG1 + FFHi**
БГШ	10,3	12,1	17,5	10,6	12,6	18,2	10,6	12,9	18,4
толпа	25,2	26,1	29,0	26,0	26,3	29,3	27,6	27,1	30,4
улица	18,0	18,9	31,4	19,5	20,3	32,7	20,2	20,9	33,4
поезд	36,1	37,2	47,4	37,8	38,6	47,9	39,4	40,2	48,6
автомобиль	45,7	50,2	53,5	46,9	50,8	53,1	46,9	51,0	53,3
среднее (ср.)	27,1	28,9	35,8	28,1	29,7	36,2	28,9	30,4	36,8
RI	–	2,5	11,9	1,5	3,6	12,6	2,6	4,6	13,4
ср. без БГШ	31,2	33,1	40,3	32,5	34,0	40,8	33,5	34,8	41,4
RI без БГШ	–	2,6	13,2	1,9	4,0	13,8	3,3	5,1	14,8

ПРИЛОЖЕНИЕ Г

Акты о внедрении

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский государственный университет аэрокосмического приборостроения»
(ГУАП)

ул. Большая Морская, д. 67, лит. А, Санкт-Петербург, 190000, Тел. (812) 710-6510, факс (812) 494-7057,
E-mail: common@aanet.ru ОГРН 1027810232680, ИНН/КПП 7812003110/783801001

_____ № _____ На № _____ от _____

УТВЕРЖДАЮ
Проректор по учебно-воспитательной работе,
доктор юридических наук, профессор
_____ / Боер В. М.
«22» декабря 2016 г.

АКТ
о внедрении результатов
диссертационной работы
Томчука Кирилла Константиновича

Комиссия, в составе: председатель – директор института радиотехники, электроники и связи профессор, д.т.н. Бестугин А. Р., члены комиссии: зам. заведующего кафедры радиотехнических систем доцент, к.т.н. Хоменко А. А., зам. заведующего кафедры по учебно-методической работе доцент, к.т.н. Бакшеева Ю. В., проф., д.т.н. Филиппов А. А., составила настоящий акт о том, что результаты исследований и практические разработки диссертационной работы Томчука Кирилла Константиновича «Сегментация речевых сигналов для задач автоматической обработки речи» используются в учебном процессе ГУАП по кафедре радиотехнических систем в следующем виде:

1. Материалы диссертационной работы используются при проведении лабораторных работ по дисциплинам «Цифровая обработка сигналов» и «Цифровая обработка сигналов и сигнальные процессоры в системах подвижной радиосвязи»;
2. Результаты исследований и разработанное программное обеспечение использовались при написании магистерских диссертаций и дипломных работ студентами групп 2420М, Z8222К, Z0222К.

Директор института радиотехники, электроники и связи, профессор, д.т.н. _____ / Бестугин А. Р.

Зам. заведующего кафедрой радиотехнических систем, доцент, к.т.н. _____ / Хоменко А. А.

Зам. заведующего кафедрой радиотехнических систем по уч.-методической работе, доцент, к.т.н. _____ / Бакшеева Ю. В.

Профессор кафедры радиотехнических систем, д.т.н. _____ / Филиппов А. А.

“УТВЕРЖДАЮ”
Технический директор
ЗАО «НПП «Иста-Системс»
Соколов В.В.
« 10 » ноября 2016 г.

А К Т

об использовании научных и практических результатов
диссертационной работы Томчука Кирилла Константиновича

Настоящим актом подтверждаем использование в компании «ЗАО «НПП «Иста-Системс» алгоритмов технологической обработки речевых сигналов, разработанных в диссертационной работе «Сегментация речевых сигналов для задач автоматической обработки речи» Томчука К. К.

Исследовательское программное обеспечение, реализующее алгоритмы временной сегментации речевых сигналов, использовалось в ОКР «Создание специализированного аппаратно-программного комплекса для автоматизации процесса выполнения комплексных исследований по материалам экстремистской направленности», шифр «Фоб», проводимой фирмой ЗАО НПП «Иста-Системс».

Разработанные материалы включают: энергетический алгоритм обнаружения речевой активности (VAD-алгоритм), алгоритм сегментации вокализованных фрагментов речевого сигнала на периоды колебаний голосовых связок (ОТ-сегментация), алгоритмы выделения основных типов речевой активности (вокализованный/шумный/взрывной), а также ряд вспомогательных алгоритмов – увеличения эффективности MFCC-параметризации при наличии шумов, обнаружения в речевом сигнале трендов и разладок, выделения огибающей.

Предложенные алгоритмы использовались в производственном процессе создания программного обеспечения специализированного аппаратно-программного комплекса для автоматизации процесса выполнения комплексных исследований по материалам экстремистской направленности.

Использование полученных Томчуком К. К. результатов позволило повысить эффективность разрабатываемых систем, существенно сократить сроки проектирования и уменьшить затраты на проведение исследовательских работ.

Директор Центра СТС
ЗАО НПП «ИСТА-Системс» _____ Сазанов В.В.

Руководитель ОКР,
ведущий программист Центра СТС
ЗАО НПП «ИСТА-Системс» _____ Лобанова М.А.