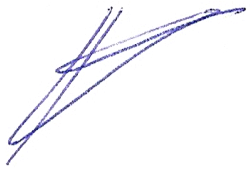
На правах рукописи



Томчук Кирилл Константинович

СЕГМЕНТАЦИЯ РЕЧЕВЫХ СИГНАЛОВ ДЛЯ ЗАДАЧ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ РЕЧИ

Специальность 05.12.13 – Системы, сети и устройства телекоммуникаций

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2017

Работа выполнена на кафедре №22 Радиотехнических систем Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет аэрокосмического приборостроения»

Научный руководитель: кандидат технических наук, с. н. с., доцент
Корнеев Юрий Алексеевич

Официальные оппоненты: **Карпов Алексей Анатольевич**
доктор технических наук, доцент, заведующий лабораторией речевых и многомодальных интерфейсов Федерального государственного бюджетного учреждения науки «Санкт-Петербургский институт информатики и автоматизации Российской академии наук» (СПИИРАН)
Тропченко Андрей Александрович
кандидат технических наук, доцент, доцент кафедры вычислительной техники Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики» (Университет ИТМО)

Ведущая организация: **Общество с ограниченной ответственностью «Центр речевых технологий» (ООО «ЦРТ»)**

Защита состоится «6» июня 2017 г. в 14:00 на заседании диссертационного совета Д 212.233.05 в Федеральном государственном автономном образовательном учреждении высшего образования «Санкт-Петербургский государственный университет аэрокосмического приборостроения» по адресу: г. Санкт-Петербург, ул. Б. Морская, 67, ауд. 53-01.

С диссертацией можно ознакомиться в библиотеке Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет аэрокосмического приборостроения» и на сайте www.guap.ru

Автореферат разослан «28» апреля 2017 г.

Ученый секретарь диссертационного совета — Овчинников Андрей Анатольевич, к. т. н., доцент

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

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 решаемой 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 технологий 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.

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. 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.

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., 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 системно-целостного 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 задачи:**

1. Провести анализ:
   а. механизмов формирования звуков речи;
   б. основных задач, возникающих при разработке алгоритмов сегментации речевых сигналов (РС);
   в. существующих подходов к сегментации РС.
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 сегментации;
   б. разработать частные алгоритмы многоуровневой сегментации РС;
   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 заключается в следующем:

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 используются 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, использовалась 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 разработанном приложении модификации темпа произнесения речи (НИР по гранту ПСП12377 правительства Санкт-Петербурга, 2012 г.; НИР по гранту МК-4934.2012.9 Президента РФ, 2012-2013 г.; НИР ПСР-3.1.2–11 по целевой программе стратегического развития ГУАП, 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: Научная сессия ГУАП (Санкт-Петербург, с 2009 по 2015); 20-я межвузовская научно-техническая конференция «Военная радиоэлектроника: опыт использования и проблемы, подготовка специалистов» (г. Санкт-Петербург, 2009); международная научная конференция «Системы и модели в информационном мире (СМИ-2009)» (г. Таганрог, 2009 г.); международная научная конференция «Современные исследовательские и образовательные технологии (СИОТ-2010)» (г. Таганрог, 2010); 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.

СОДЕРЖАНИЕ РАБОТЫ

Во введении представлена общая характеристика диссертационной работы: обоснована актуальность проблемы, сформулированы цели и задачи исследования, определена научная новизна и практическая значимость.

В первом разделе осуществлен аналитический обзор публикаций, посвященных тематикам речеобразования и речевосприятия, автоматическому анализу речевых сигналов в целом и автоматической временной сегментации в частности.

О механизмах восприятия речи человеком достоверно известно немногое, так как изучение процесса обработки человеческим мозгом получаемой информации является крайне сложной задачей. Как следствие, существующие алгоритмы обработки речевых сигналов имеют мало общего с принципами человеческого речевосприятия.

Временная сегментация речевых сигналов является базовой задачей в любой голосовой системе и необходима для ее эффективной работы. 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), на отдельные периоды колебаний голосовых связок в огласованных звуках.

Речевое сообщение является дискретным и может быть представлено как последовательность фонем, в русском языке принято выделять 43 фонемы (37 согласных и 6 гласных). Несмотря на глубокую связь произносимых фонем с их сигнальными параметрами, в технической литературе отсутствует единство в вопросе структурной классификации фонем и их вариаций.

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. Вопрос выбора набора сигнальных параметров для построения алгоритма сегментации является нетривиальным и представляет большую сложность и важность. Отсюда также следует большое разнообразие подходов к решению отдельных задач сегментации, основные из которых проанализированы в рамках первого раздела.

На результатах работы алгоритмов сегментации речевых сигналов основывается работа широкого класса речевых приложений, выполняющих определенную конечную задачу речевых технологий.

Наиболее широко применяемыми в самостоятельном виде алгоритмами сегментации являются VAD-детектор (определение временных границ речевой активности) и алгоритм выделения границ периодов основного тона (ОТ; выделение отдельных колебаний голосовых связок) – различные реализации таких алгоритмов демонстрируют огромное количество возможных подходов к решению задач сегментации.

Второй раздел посвящен исследованию сигнальных особенностей реализаций фонем на примере русского языка.

После изучения фонетического строения русской речи принято решение о рассмотрении помимо основных 43 фонем русского языка также ряда существенных аллофонов, определяющих звучание фонем в зависимости от положения в потоке речи: длительные согласные ([д:], [ж:], [н:]), аллофоны гласных [и^{е}], [и^{э}], [ы^{э}], редуцированные гласные [^] (преимущественно первая слабая позиция для гласных *А, О, Э*) и [ь], [ъ] (вторая слабая позиция гласных *А, О, Э, Е, Я* в положении соответственно после мягких и твердых согласных).

В рамках исследования поставлена цель обеспечения появления каждого аллофона у каждого диктора и в определенной позиции (начало, середина и конец слова) не менее двух раз. Для достижения приемлемой оценки «дикторозависимости» параметров все слова произносились двумя дикторами: мужского и женского пола. Исходя из приведенных условий для проведения исследования, подготовлен перечень из 184 слов и их транскрипций. Для автоматизации транскрибирования создан алгоритм обработки текста, учитывающий основные правила произношения и использующий словарь ударений русских слов.

С помощью специально разработанного программного обеспечения произведена ручная сегментация записанных для исследования фонограмм. В результате сформирована база данных, содержащая основную информацию о более чем 2000 реализациях русских фонем: временную позицию относительно начала слова; набор сэмплов звука; соответствующий аллофону фонетический символ; метаданные о дикторе; порядковую позицию фонемы относительно начала слова, конца слова, ударной гласной.

Для всех имеющихся реализаций фонем база данных дополнена результатами вычисления основных сигнальных параметров: длительность звука; средняя мощность; нормированная сумма модулей отсчетов; энергия; частота переходов через нуль; мел-частотные кепстральные коэффициенты MFCC (вектор значений); количество переколебаний на периоде основного тона (только для вокализованных). Для приведенных в списке энергетических параметров важно упомянуть о едином сеансе записи фонограмм и их общей нормализации к уровню *±1*.

Использование базы данных позволяет выполнять многокритериальное извлечение статистических данных: по группе фонем, по диктору, по признаку ударности, по положению фонем относительно границ слова, других фонем, ударного гласного – и предоставляет важную количественную информацию, необходимую для разработки и установки пороговых значений в алгоритмах сегментации речевых сигналов. Такая база является дополняемой как по количеству реализаций фонем, так и по перечню вычисляемых параметров.

С учетом базовых положений фонетики и результатов проведенного анализа, сформирована иерархическая структура групп (таксономия) исследуемых аллофонов русской речи, отражающая основные уровни сегментации и нюансы эффективности VAD и OT-сегментации (рисунок 1).

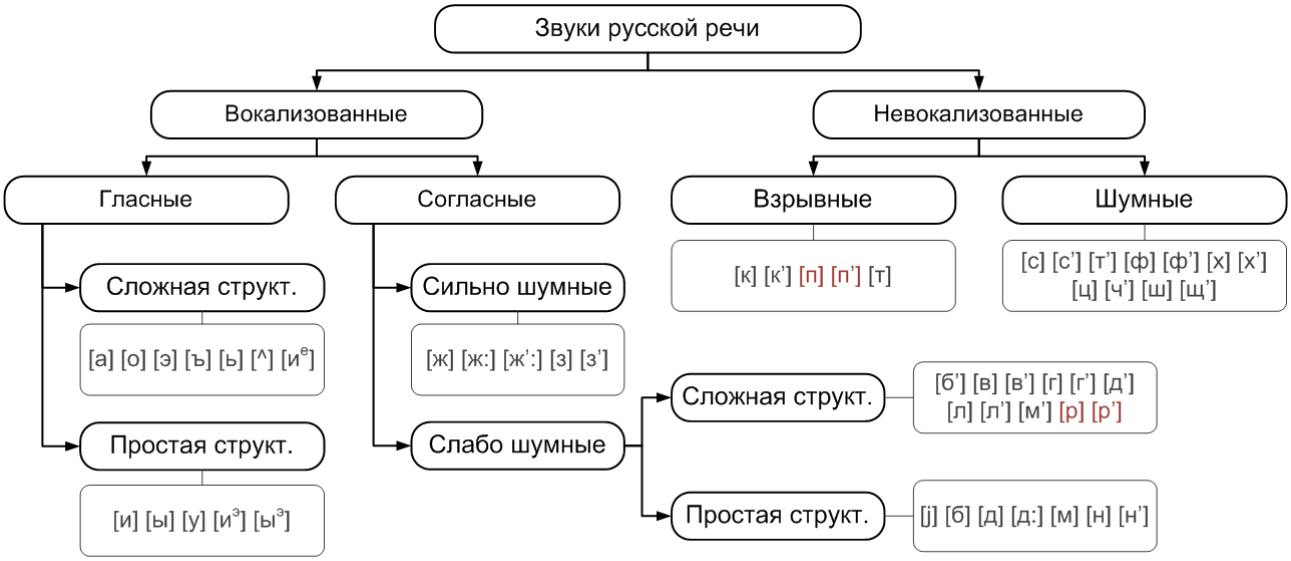


Рисунок 1. Таксономия русских звуков для задачи сегментации

Для гласных звуков, в особенности, ударных или находящихся в первой слабой позиции, характерна относительно высокая средняя мощность. Разделение вокализованных гласных и согласных на группы «сложной» и «простой структуры» основывается на количестве переколебаний в структуре низкочастотной компоненты периода основного тона и проявляется в чистом виде для голосов с высоким тоном (т.е. женских). Сложная структура периода основного тона является значимой причиной снижения надежности алгоритмов ОТ-сегментации.

У вокализованных согласных форманты выше второй значительно менее выражены, чем у гласных и, порой, трудноразличимы. В данной группе присутствуют звуки со значительной шумовой составляющей, снижающей точность VAD-сегментации: это щелевые звонкие [ж], [ж:], [ж’:], [з] и [з’], 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.

Для группы невокализованных взрывных характерна малая длительность звука (в среднем меньше длительности паузы-смычки), а также наличие одного-двух «взрывов», вызываемых резким раскрытием препятствия в речевом тракте и выражающихся в кратковременном повышении интенсивности сигнала по всему речевому диапазону частот.

К невокализованным шумным относятся аффрикаты и глухие спиранты. Этой группе звуков характерна большая, нежели у невокализованных взрывных, длительность. Палатализованный [т’] отнесен к подгруппе шумных в силу сравнительно большой длительности и отсутствия характерного для других взрывных звуков резкого амплитудного скачка.

Таким образом, в зависимости от конфигурации речевого аппарата, при произнесении тех или иных звуков в речевом сигнале наблюдаются характерные особенности, обуславливающие возможные уровни временной сегментации. Полученные в рамках второго раздела статистические значения параметров звуков, а также разработанная таксономия позволяют реализовать алгоритмы временной сегментации речевых сигналов с отнесением выделяемых сегментов к конкретным группам звуков.

В третьем разделе описывается разработка частных алгоритмов сегментации речевых сигналов и ряда вспомогательных алгоритмов. На основе информации о строении речевого сигнала, 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 2). Следует отметить, что разбиение на периоды основного тона необязательно является финальной стадией сегментации: в дальнейшем на ее основе может быть выполнено разбиение вокализованных фрагментов на отдельные аллофоны / группы из однотипных аллофонов.

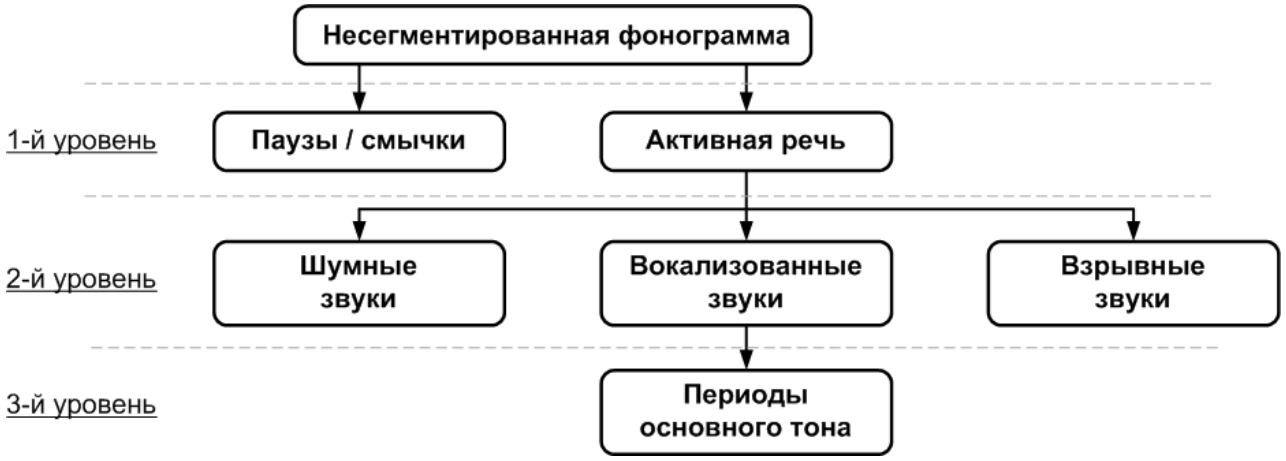


Рисунок 2. Три уровня сегментации речевого сигнала

Для сегментации на все представленные на рисунке 2 уровни в рамках третьего раздела диссертации разработаны соответствующие алгоритмы.

Разработан алгоритм выделения огибающей речевого сигнала, который, как будет показано далее, лег в основу алгоритма ОТ-сегментации вокализованной речевой активности. Непосредственной задачей алгоритма является отбор необходимых отсчетов РС, совокупность которых в итоге образует огибающую.

Для уменьшения вычислительной сложности на первом шаге производится выделение множества *M* локальных максимумов РС:

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

где *i* – номера отсчетов исходного РС, *x(i)* – значения отсчетов исходного РС.

Далее номера отсчетов, составляющих огибающую, итеративно определяются из элементов множества *M*:

$$\begin{cases} K = \{m \in M \mid a_n < m \le a_n + L\} \\ a_{n+1} \in K \\ \forall k \in K \;\; p(a_{n+1}) \ge p(k) \\ \{k \in K \mid p(a_{n+1}) = p(k) \wedge a_{n+1} < k\} = \varnothing \end{cases}, \tag{2}$$

где a_n – текущий номер отсчета огибающей; a_{n+1} – следующий номер отсчета огибающей; смысл функции $p(k)$ состоит в определении наклона линии, проведенной на осциллограмме сигнала через точки (a_n, $x(a_n)$) и (k, $x(k)$):

$$p(k) = \frac{x(k) - x(a_n)}{k - a_n}. \tag{3}$$

Параметром L регулируется расстояние в отсчетах между точками огибающей, которое алгоритм стремится получить, не превышая его.

На основе данного алгоритма выделения огибающей разработан алгоритм ОТ-сегментации по максимумам отдельных периодов ОТ. Для этого в алгоритм выделения огибающей помимо параметра L, ограничивающего в данном случае максимально возможный период ОТ человеческого голоса, вводится второй схожий параметр, ограничивающий минимально возможный период ОТ. Кроме того, кандидаты из множества K проходят ряд дополнительных условий, направленных на борьбу с пропусками границ ОТ-сегментов.

Эффективность разработанного алгоритма (далее обозначен как алгоритм EOT) проанализирована в сравнении с рядом других известных современных алгоритмов ОТ-сегментации: DYPSA, SEDREAMS и SE-VQ. Экспериментальное исследование произведено на фонограммах базы TIDIGITS с использованием в качестве вспомогательных алгоритмов оценки частоты ОТ PEFAC и RAPT, а также алгоритма SRH оценки степени вокализованности фрагментов РС. Тестовые РС искажены естественными помехами 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 рисунке 3 приведены графики полученных оценок показателей надежности алгоритмов ОТ сегментации: *IDR* (частота правильного определения сегментов), *MR* (частота пропуска сегментов) и *FAR* (частота сегментов с ложными обнаруженными границами). По данным показателям разработанный алгоритм EOT показывает значения лучше, чем у 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.

Также на рисунке 3 показаны значения, полученные для показателя точности ОТ-сегментации: *IDA* (стандартное отклонение ошибки определения временной границы периода ОТ). Разработанный алгоритм не выходит за рамки значений *IDA* сторонних алгоритмов при ОСШ не менее 5 дБ.

На рисунке 4 приведены результаты оценки показателя скорости работы алгоритмов *SF*, определяемого как отношение времени обработки всех анализируемых фонограмм к их общей длительности. По данному показателю разработанный алгоритм многократно превосходит аналоги.

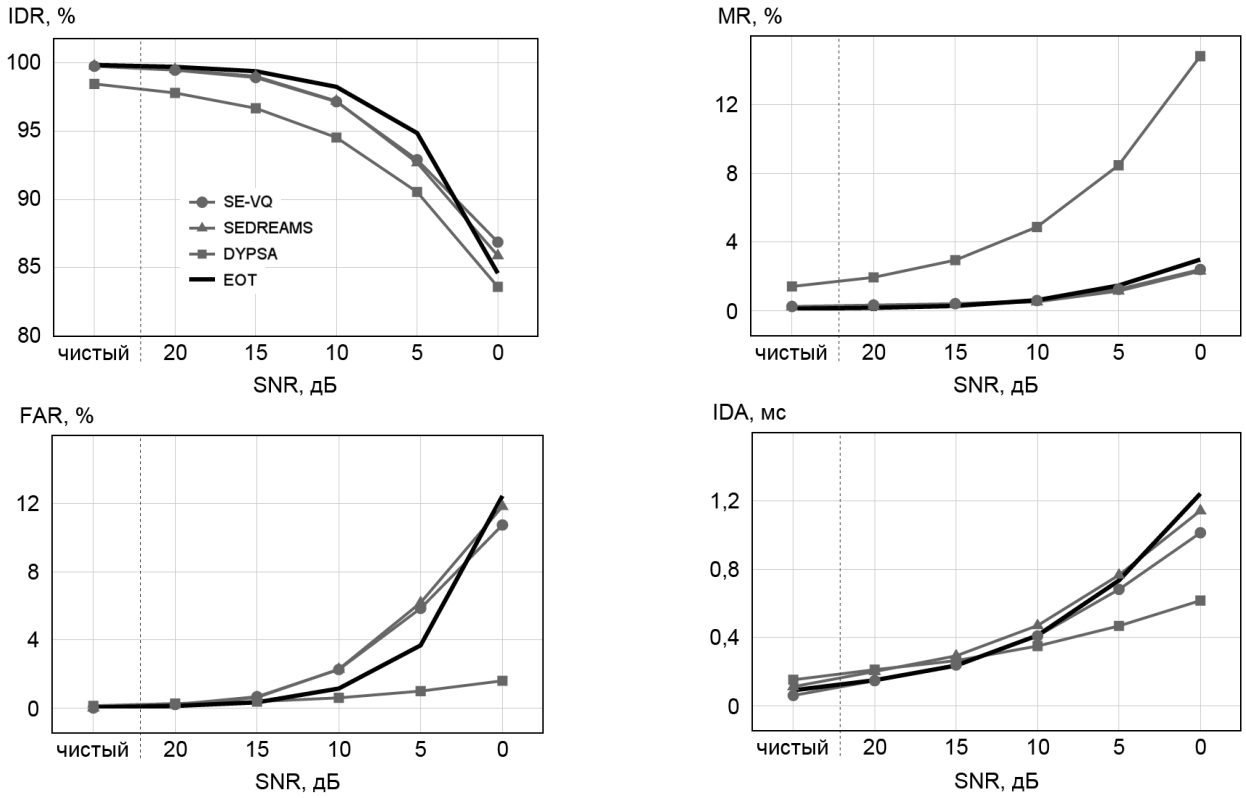


Рисунок 3 – Показатели надежности (IDR, MR, FAR) и точности (IDA) алгоритмов ОТ-сегментации

Далее в третьем разделе предложен метод повышения эффективности параметризации РС MFCC-коэффициентами. Данные коэффициенты широко используются при решении речевых задач, включая сегментацию, однако, имеют крайне низкую шумоустойчивость.

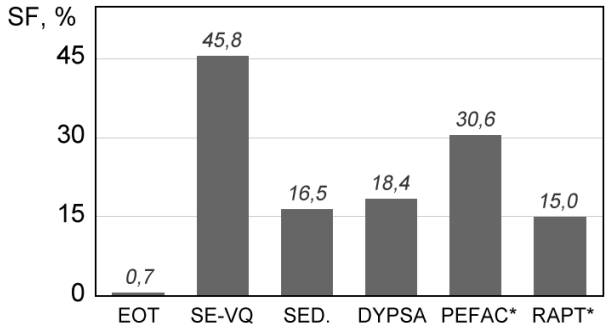


Рисунок 4 – Показатель скорости работы алгоритмов ОТ-сегментации

Предложенный метод основан на двух гипотезах:

а. о положительном влиянии применения к РС психоакустической модели (ПАМ) восприятия звукового анализатора человека;
б. о возможности воздействия на спектр сигнала на частотах высших гармоник ОТ с целью получения более выраженной формантной картины.

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. Для воздействия на значения спектра мощности на частотах, кратных частоте ОТ, предложено следующее преобразование спектральной плотности мощности (далее упоминается как алгоритм ОТ):

$$P(k)=\begin{cases}(1+0,5|i|)P_x(k)Kp_v, & k=\left[nN\dfrac{f_{OT}}{F_S}\right]+i, n\in\mathbb{Z}, n<\dfrac{F_S}{2f_{OT}}, i=\overline{-1,1}\\ P_x(k) & k\neq\left[nN\dfrac{f_{OT}}{F_S}\right]+i, n\in\mathbb{Z}, n<\dfrac{F_S}{2f_{OT}}, i=\overline{-1,1}\end{cases}, \quad (4)$$

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 подобрано значение 1,3; квадратными скобками показана операция округления до ближайшего целого.

Внедрение ПАМ эффекта слуховой маскировки является известным методом повышения эффективности MFCC-параметризации. Для этих целей используется механизм латерального торможения, эффективно описывающий природу одновременного (частотного) маскирования, и реализуемый путем фильтрации спектра мощности РС.

В диссертационной работе предложено применить другой подход к учету одновременного маскирования, широко используемый 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.

В диссертационной работе исследование эффективности применения ПАМ вычисления глобального маскирующего порога производится в сравнении с ПАМ, основанной на механизме латерального торможения. В качестве первой используется алгоритм, применяемый в стандарте сжатия аудиосигналов ISO/IEC MPEG-1 Layer 1 (далее – алгоритм MPEG1). В качестве второй ПАМ рассматривается подалгоритм LI (Lateral Inhibition, латеральное торможение), входящий в состав алгоритма LTFC – модифицированного алгоритма MFCC-параметризации.

Экспериментальное исследование результатов модификации производилось на базе фонограмм TIDIGITS. В качестве критерия оценки эффективности предложенных решений использовалась частота распознавания в системе распознавания одиночных слов (наиболее распространенный подход исследования эффективности параметризации в дикторонезависимых системах), в которой для параметризации РС использованы только вектора MFCC-коэффициентов.

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.

В таблице 1 представлены полученные пословные точности распознавания, а также значения относительного улучшения результатов распознавания с усреднением по перечисленным выше шумам в сравнении с традиционным алгоритмом MFCC. Относительное улучшение *RI* рассчитывается по формуле:

$$RI = \frac{RR_A - RR_{MFCC}}{100 - RR_{MFCC}} \times 100\%, \qquad (5)$$

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

Предложенное преобразование спектра мощности РС, заключающееся в усилении спектральных составляющих на частотах, кратных частоте ОТ, при низких ОСШ может быть эффективно использовано совместно с психоакустическими модификациями LI и MPEG1: строки LI+OT, MPEG1+OT в таблице 1.

Таблица 1. Пословная точность распознавания (%) и относительные улучшения *RI* (%, правый столбец) в сравнении с алгоритмом MFCC

Алгоритм\ОСШ	20 дБ	15 дБ	10 дБ	5 дБ	0 дБ	средн. 0-20 дБ	средн. *RI* 0-20 дБ
MFCC(13)	75,7	68,4	58,4	44,6	31,2	55,7	–
LI	75,0	68,4	59,4	46,6	33,1	56,5	1,2
MPEG1	75,8	71,9	65,1	55,0	40,3	61,6	11,9
OT	75,8	68,9	59,0	46,0	32,5	56,4	1,6
LI+OT	75,1	68,3	59,7	47,3	34,0	56,9	1,8
MPEG1+OT	75,7	71,9	65,1	55,3	40,8	61,8	12,1

В рамках третьего раздела приводятся также следующие разработанные алгоритмы сегментации и вспомогательные алгоритмы:

1. Энергетический VAD-алгоритм (1-й уровень сегментации). Предложенная реализация алгоритма характерна высокой частотой среза ФНЧ (превышает частоту ОТ большинства людей), введением операции логарифмирования (для увеличения разрешающей способности гистограммы, по которой формируется адаптивный порог, в области малых значений и приближения логики работы алгоритма к механизму звуковосприятия человека – закон Вебера-Фехнера). При принятии решений о границах сегментов учитываются, в том числе, граничные значения длительностей смычек и отдельно стоящих звуков, полученные в рамках второго раздела работы.
2. Сегментация «шумный/нешумный». В разработанном алгоритме выделения из сегментов речевой активности шумных звуков параметром для принятия решения является отношение стандартного отклонения отсчетов РС, пропущенного через ФНЧ, призванный сгладить выбросы сигнала в шумных фрагментах, к стандартному отклонению отсчетов исходного РС.
3. Сегментация «вокализованный/невокализованный» на основе анализа локальных экстремумов АКФ фрагмента РС.
4. Сегментация на периоды ОТ корреляционным методом с коррекцией до точки ближайшего пересечения нуля в направлении снизу вверх пропущенного через ФНЧ РС.
5. Обобщенный подход к сегментации, дающий возможность осуществления сегментации РС требуемого уровня (до широких фонетических классов) и использующий сформированную в разделе 2 базу данных параметров исследованных аллофонов русского языка.
6. Выявление переходных участков вокализованных сегментов по структурным и амплитудно-структурным метрикам разладок НЧ-структур периодов ОТ.
7. Отсечение остаточных колебаний голосовых связок при переходе от вокализованного к паузе по энергетическому критерию.

Четвертый раздел работы посвящен анализу особенностей применения алгоритмов сегментации в речевых приложениях. Освещены вопросы необходимости различных уровней сегментации для задач сжатия РС, распознавания команд, идентификации и верификации диктора, шумоподавления в РС. Детально применение результатов автоматической временной сегментации рассмотрено на примере разработанного алгоритма модификации темпа произнесения речевой фонограммы.

Операция изменения темпа произнесения позволяет так обработать РС, чтобы скорость произнесения изменилась в заданное количество раз, но при этом тембр (частота ОТ) голоса диктора оставался неизменным.

Традиционно для модификации характеристик речи используется вокодерный подход «анализ-синтез» с изменением требуемым образом полученных на этапе анализа оценок спектральных параметров РС.

В разработанном в рамках диссертационной работы алгоритме модификации темпа речи для решения задачи предлагается использовать результаты автоматической многоуровневой временной сегментации РС. Такой подход дает возможность для разных типов сегментов применять разные подалгоритмы обработки, а также выполнять ускорение/замедление с разными парциальными коэффициентами.

Важно отметить, что в естественной речи при разных темпах произнесения одной и той же фразы длительности разных типов фонем меняются в разной степени. Наиболее сильным изменениям подвержены гласные, а при очень быстром темпе речи отдельные гласные могут исчезать. 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 в той же степени, что и участки речи.

Для изменения темпа речи предложенным способом, необходимо выполнить сегментацию РС на паузы, вокализованные, шумные и взрывные, а для вокализованных – выполнить ОТ-сегментацию.

На вход алгоритма модификации темпа речи оператором вводится интегральный коэффициент растяжения K фонограммы в целом (для ускорения темпа произнесения коэффициент K устанавливается в значения <1).

По заданному 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. Взрывные звуки при изменении темпа следует оставлять без изменений: $K_{вз}$=1.

Особое значение имеет изменение темпа за счет вокализованных сегментов: при растяжении производится добавление, а при сжатии – замещение нескольких периодов ОТ одним синтезированным. Для лучшего качества звучания для синтеза применяется векторная интерполяция двух периодов ОТ исходной фонограммы.

Пример участка фонограммы, являющейся результатом модификации темпа речи, показан на рисунке 5.

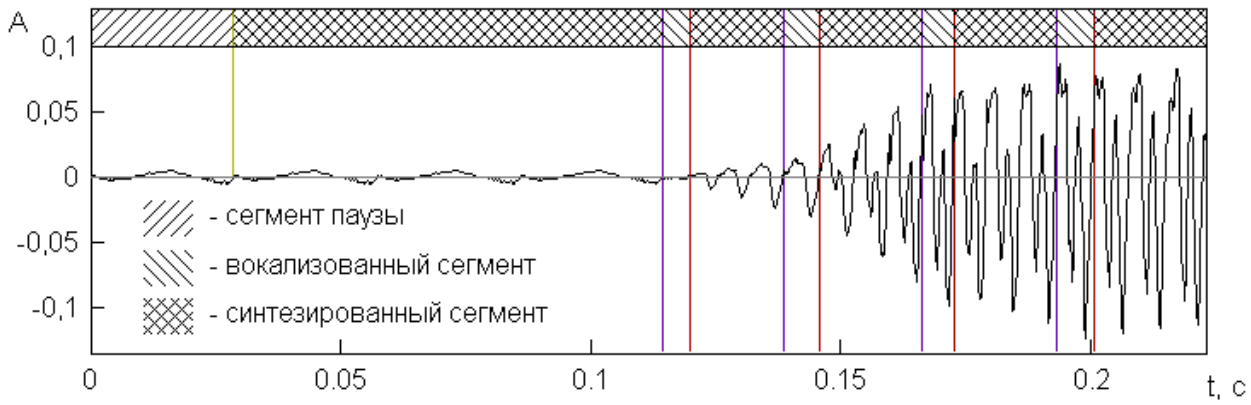


Рисунок 5 – Участок фонограммы с замедленным темпом произнесения

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

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 разного качества и содержащих в общей сложности по два мужских и женских голоса. Каждая тестовая фонограмма была модифицирована сравниваемыми алгоритмами с четырьмя коэффициентами изменения темпа: ускорение в 1,5 и 2 раза; замедление в 2 и в 3 раза.

По оценкам экспертов сформированы сводные таблицы: результаты ранжирования в режиме замедления (таблица 2) и ускорения (таблица 3) темпа, – в обоих случаях разработанный алгоритм занимает лидирующую позицию.

Таблица 2 – 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 3 – 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 количественной оценки степени согласованности мнений экспертов вычислен дисперсионный коэффициент множественной конкордации Кендалла. Данный коэффициент W может принимать значения от нуля до единицы, и при значениях $W > 0{,}5$ оценки экспертов считаются в достаточной мере согласованными. Вычисленный по таблицам 2 и 3 коэффициент конкордации в режиме замедления темпа речи составляет $W = 0{,}89$, в режиме ускорения – $W = 0{,}63$. При ускорении темпа различия между фонограммами, полученными разными алгоритмами, достаточно сложно выявить на слух, чем можно объяснить уменьшение значения W по сравнению с режимом замедления.

Высокие оценки разработанного алгоритма в режиме замедления темпа связаны с отсутствием типового для других аналогичных алгоритмов артефакта звучания, вызываемого спектральными искажениями речи и напоминающего на слух эффект реверберации. В режиме ускорения темпа значимым преимуществом разработанного алгоритма является различие парциальных коэффициентов изменения отдельных сегментов: в отличие от других алгоритмов, короткие невокализованные звуки не ужимаются по времени до полной редукции и остаются отчетливо различимыми на слух.

Таким образом, предложенное решение задачи модификации темпа речи демонстрирует высокий уровень разработанных в рамках третьего раздела алгоритмов автоматической временной сегментации РС, а также эффективность применения сегментационного подхода.

В заключении приведены основные результаты исследования и сформулированы выводы.

В приложениях представлена методика исследования сигнальных особенностей звуков; вычисленные оценки основных сигнальных параметров звуков; пословные точности распознавания и относительные улучшения, полученные применением различных модификаций алгоритма MFCC-параметризации.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. В диссертационной работе 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, отражающая их важные с точки зрения задачи временной сегментации сигнальные особенности.

3. Создан программный комплекс, объединяющий файловое хранилище цифровых фонограмм, хранилище базы данных, функционал для обработки речевых сигналов, текстов, систематизации данных. Комплекс позволяет производить, в частности, автоматизированное транскрибирование русских слов, ручную сегментацию фонограмм, вычисление сигнальных параметров звуков, формирование выборок вычисленных значений параметров для произвольных групп аллофонов / типов сегментов речевой активности.
4. Разработаны и исследованы частные алгоритмы временной сегментации речевого сигнала на различные уровни, а также ряд вспомогательных алгоритмов.
5. Разработан алгоритм модификации в широком диапазоне темпа произнесения речевого сигнала, основанный на раздельной обработке основных типов сегментов речевого сигнала. Алгоритм при сравнении с существующими аналогами показал низкое количество артефактов звучания модифицированного сигнала.

СПИСОК ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Томчук К. К., Зилинберг А. Ю. Разработка и исследование высококачественного алгоритма модификации темпа произнесения речи // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2009. – С. 79–82.
2. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Разметка фонограммы на периоды основного тона методами структурного анализа речевого сигнала // Сборник докладов 20-й межвузовской научно-технической конференции «Военная радиоэлектроника: опыт использования и проблемы, подготовка специалистов». – СПб. : ВМИРЭ, 2009.
3. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Высококачественный алгоритм модификации темпа произнесения речи: разработка и апробация // Международная научная конференция «Системы и модели в информационном мире (СМИ-2009)»: Материалы конференции. – Таганрог: ТТИ ЮФУ (ТРТУ), 2009. – С. 80–91.
4. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Разработка и исследование алгоритмов многоуровневой временной сегментации речевых сигналов // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2010. – С. 21–25.
5. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Анализ трендов и разладок структуры ОТ-кластеров вокализованных сегментов речи // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2010. – С. 70–73.
6. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Особенности сегментации речевых сигналов для задач автоматической обработки речи // Международная научная конференция «Современные исследовательские и образовательные технологии (СИОТ-2010)»: Материалы конференции. – Таганрог: ТТИ ЮФУ (ТРТУ), 2010. – С. 43–54.
7. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Разработка алгоритмов подавления импульсных помех в трактах передачи речевых сигналов // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2011. – С. 20–23.

8. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Анализ характеристик импульсных помех в тракте передачи речевых сигналов // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2011. – С. 19–20.
9. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Многоуровневая временная сегментация речевых сигналов в задаче модификации темпа воспроизведения фонограмм // Вопросы оборонной техники; Серия 16. Технические средства противодействия терроризму. – М. : НТЦ «Информтехника», 2011. – С. 85–93 **(из перечня журналов ВАК)**.
10. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Обзор технологических алгоритмов обработки речевых сигналов для задач идентификации и верификации диктора // Материалы всероссийской научной конференции «Перспективы развития гуманитарных и технических систем». – Таганрог: Изд-во ТТИ ЮФУ, 2011. – Ч. 2 – С. 62–74.
11. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Системный подход к сегментации и параметризации речевых сигналов // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2012. – С. 53–56.
12. Томчук К. К., Зилинберг А. Ю. Общая методика сравнения эффективности алгоритмов сегментации и параметризации речевых сигналов // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2012. – С. 56–59.
13. Томчук К. К., Зилинберг А. Ю. Иерархическая модель и базовые алгоритмы временной сегментации речевых сигналов // Ученые записки Петрозаводского государственного университета. – Петрозаводск: ПетрГУ, 2013. – №2(131). – С. 114–119 **(из перечня журналов ВАК)**.
14. Томчук К. К., Зилинберг А. Ю., Корнеев Ю. А. Системные вопросы многоуровневой временной сегментации речевых сигналов // Сборник докладов Научной сессии ГУАП. – СПб. : ГУАП, 2013. – С. 80–83.
15. Томчук К. К. Применение частотного маскирования при MFCC-параметризации речи на фоне шумов // Информационно-управляющие системы. – Санкт-Петербург, 2016. – №3(82). – С. 8–14 **(из перечня журналов ВАК)**.

Авторские свидетельства

16. Томчук К. К. Система модификации темпа произнесения речевых фонограмм // Свидетельство о регистрации электронного ресурса № 20862 от 17.04.2015, регистрационный номер ВНТИЦ 50201550159.