Федеральное государственное автономное образовательное
учреждение высшего образования
«Российский университет дружбы народов»

УДК: 681.5

На правах рукописи

БЕССОНОВ Максим Александрович

Алгоритмы интерпретации просодических признаков речи при обработке аудиосообщений

Специальность 05.13.15 –
Вычислительные машины, комплексы и компьютерные сети

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2017

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Российский университет дружбы народов»

Научный руководитель	**Шалимов Игорь Анатольевич** доктор технических наук, профессор сотрудник в/ч 33965, г. Москва
Официальные оппоненты:	**Орлов Юрий Николаевич** доктор физико-математических наук, старший научный сотрудник, заведующий отделом Федерального государственного учреждения "Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук", г. Москва
	Мельников Сергей Юрьевич кандидат физико-математических наук, заместитель генерального директора ООО «Лингвистические и информационные технологии», г.Москва
Ведущая организация	Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г.Москва

Защита диссертации состоится 18 декабря 2017 г. в 11 часов 00 мин. на заседании диссертационного совета Д002.226.03 при Институте проблем управления им. В.А. Трапезникова РАН по адресу: 117997, Москва, ул. Профсоюзная, д. 65, Малый конференц-зал.

С диссертацией можно ознакомиться в Научной библиотеке Федеральном государственном учреждении науки Институт проблем управления им В.А Трапезникова РАН.

(Отзывы на автореферат просьба направлять по указанному адресу.)

Автореферат разослан «____» _____ 2017 г.

Ученый секретарь
диссертационного совета № 3. Д002.226.03

кандидат технических наук Кулинич Александр Алексеевич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В настоящее время вычислительные комплексы присутствуют во всех областях человеческой деятельности, осуществляя обработку большого количества разнородной информации. Одним из видов такой информации является речевая информация – аудиосообщения, передаваемые по компьютерным сетям, то есть IP-телефония, либо по сетям телефонной и радиосвязи. Вычислительные комплексы и машины решают следующие задачи по обработке речи – идентификацию и верификацию диктора, определение языка аудиосообщения, синтез речи, перевод речи в текст, конвертацию форматов представления речевых данных. При этом эти задачи могут решаться либо в режиме диалога человека и ЭВМ, либо в датацентрах при обработке информации. Такие вычислительные комплексы должны иметь соответствующие алгоритмы обработки, ввода-вывода речевой информации.

Поскольку речь передается по каналам связи, она подвергается различным преобразованиям. Если канал связи имеет узкую полосу пропускания, то применяются низкоскоростные кодеки речи и вокодеры. При такой обработке речи из нее удаляется значительная часть информации, которая может быть получена из чистой речи. Результатом работы вокодеров являются наборы параметров, в которых всегда присутствуют такие акустические признаки речевого сигнала, как частота основного тона и усиление для текущего квазистационарного сегмента, а также параметр тон-шум. Ввиду значительного редуцирования информации системы автоматического определения языка, основанные на различных алгоритмах вычисления акустических параметров, перестают определять язык с заданной достоверностью. В связи с этим задача определения языка аудиосообщения на основе параметров, вычисляемых низкоскоростными вокодерами, без восстановления исходной формы речевого сигнала является актуальной.

Решение задачи определения языка аудиосообщения лежит в области лингвистики и математики. Существуют несколько подходов к определению языка аудиосообщения, которые реализуются в системах автоматического определения языка.

Практическая реализация того или иного подхода основана на использовании какого-либо математического аппарата (решающего правила) и словаря признаков, в качестве которых могут быть использованы акустические параметры, выделяемые на коротких сегментах, последовательности фонем, просодические признаки речи человека, комбинации больших групп фонем, которые могут составлять слова.

Достоинством просодических признаков является то, что их акустической основой являются частота основного тона и кратковременная энергия речевого сигнала, и как раз эти параметры передаются в системах связи, работающих на вокодерах. В связи с этим отсутствует необходимость восстанавливать исходную форму речевого сигнала. В то же время восстановление исходной формы речевого сигнала необходимо для работы систем на основе акустического, фонотактического и лексического подходов.

В случае, если достоверность определения языка по речи, подвергнутой вокодерной обработке, будет недостаточна для решения поставленной задачи, либо в канале связи будет присутствовать речь, обработанная гибридным вокодером, то возможна реализация смешанной системы, в которой на первом этапе аудиосообщение будет отнесено к какой-либо группе языков, а на втором этапе определение языка будет реализовываться не на всей базе данных, а только внутри этой группы. Для реализации первого этапа предлагается использовать просодические признаки речи человека, описываемые широкими фонетическими категориями. Реализация второго этапа предполагает восстановление исходной формы речевого сигнала и применение акустического, фонотактического либо лексического подходов.

Цель работы заключается в разработке новых эффективных алгоритмов интерпретации просодических признаков речи и методики их использования при решении задач обработки аудиосообщений.

Задачи исследования. Для достижения поставленной цели в работе поставлены и решены следующие задачи:

1. анализ существующих подходов определения языка аудиосообщения;
2. анализ различий между языками на просодическом уровне;
3. анализ алгоритмов работы низкоскоростных вокодеров;
4. разработка способов описания просодических признаков речи диктора;
5. разработка алгоритмов интерпретации просодических признаков речи;
6. выбор математического аппарата для классификации языков по просодическим признакам;
7. разработка методики использования алгоритмов интерпретации просодических признаков;
8. экспериментальная оценка алгоритмов.

Основные положения, выносимые на защиту состоят в следующем:

1. алгоритм интерпретации просодических признаков на основе широких фонетических категорий;

2. алгоритм интерпретации просодических признаков на основе кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий;

3. методика использования алгоритмов интерпретации просодических признаков речи в задачах определения языка аудиосообщения, в том числе без восстановления исходной формы речевого сигнала;

4. результаты оценки эффективности разработанных алгоритмов.

Научная новизна – заключается в разработке новых научно-обоснованных алгоритмов, комплексно описывающих просодические признаки речи диктора на основе широких фонетических категорий и на основе кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий и методики применения разработанных алгоритмов в задачах определения языка аудиосообщения, в том числе без восстановления исходной формы речевого сигнала в условиях передачи речи по системам вокодерной связи.

Практическая значимость. Заключается в широком спектре сфер практического применения результатов:

- разработанные алгоритмы пригодны для использования в коммерческих организациях и государственных компаниях, занимающихся вопросами специальной обработки данных, созданием вычислительных комплексов анализа речи, в том числе, передаваемой по сетям связи низкого качества, а также речевой аналитикой, а также любых других вычислительных систем, предназначенных для предоставления услуг персональным пользователям средствами голосового управления;
- исследование подходов определения языка аудиосообщения, отличий языков на просодическом уровне делает работу ценной для учебного процесса в ВУЗах, имеющих потоки подготовки по специальностям, связанным с построением или эксплуатацией вычислительных комплексов, обработкой информации, речевой аналитикой средствами вычислительных машин.

Кроме того, практическая значимость подтверждается актами внедрения результатов диссертационного исследования.

Объект исследования – системы определения языка аудиосообщения, реализуемыев вычислительных комплексах.

Предмет исследования – алгоритмы интерпретации просодических признаков речи и методики их применения в задачах специальной обработки аудиосообщений в вычислительных комплексах.

Научная задача – на основе анализа существующих подходов к определению языка аудиосообщения и просодических различий языков разработать эффективные алгоритмы интерпретации просодических признаков речи для их применения при специальной обработке аудиосообщений, подвергнутых вокодерному преобразованию, в том числе без восстановления исходной формы речевого сигнала.

Методы исследования. В работе использовались методы распознавания образов, методы статистического анализа, методы корреляционного анализа, математическое моделирование, компьютерное моделирование, методы обработки экспериментальных данных, методы цифровой обработки сигналов.

Личный вклад – автором лично получены теоретические и практические результаты работы, в случае заимствования материалов приведены их источники, автором лично проведены теоретические исследования, разработаны и реализованы алгоритмы, проведены компьютерные эксперименты, подготовлены публикации по диссертационной работе.

Достоверность. Результаты диссертационной работы обоснованы с использованием методов прикладной лингвистики, применением искуственных нейронных сетей, известных теоретических сведений о строе языков, проверены экспериментально, а также подтверждаются актами внедрения в коммерческие структуры и учебный процесс.

Реализация результатов работы. Результаты диссертационной работы использованы при выполнении ЗАО «ПАСИТ» ряда работ, а именно: ОКР шифр «Кристалл-13», Государственный контракт от 01 марта 2013 года № ЕГО-051-13; ОКР шифр «Клиент», Государственный контракт от 08 апреля 2013 года № КГО-001-13; ОКР шифр «Штурман-П», Государственный контракт от 01 февраля 2014 года № 735/ЕГО/Р/2014.

Также полученные результаты диссертационной работы внедрены в учебный процесс – лабораторные и практические занятия по дисциплине «Цифровая обработка сигналов» кафедры «Управление и защита информации» в ФГБОУ ВО «Московский государственный университет путей сообщения императора Николая II» (МГУПС (МИИТ).

Апробация результатов работы. Результаты диссертационного исследования были апробированы на XI Международной научно-технической конференции «Физика и радиоэлектроника в медицине и

экологии» (ФРЭМЭ'2014) (1 - 3 июля 2014 года, ВлГУ, г. Владимир), на научно-практических конференциях в Академии ФСО России в 2011, 2013, 2017 годах, на семинарах в Институте проблем управления РАН (г. Москва).

Публикации. По материалам диссертации опубликовано 6 работ, 5 из которых в изданиях, входящих в Перечень ведущих рецензируемых научных журналов и изданий, формируемый Высшей аттестационной комиссией.

Структура и объем работы. Диссертация содержит введение, четыре главы, заключение, список литературы, 3 приложения.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационного исследования, сформулированы цель и основные задачи, решение которых необходимо для достижения поставленной цели. Кратко изложено основное содержание работы.

В первой главе приведена классификация систем определения языка аудиосообщения, рассмотрены требования, предъявляемые к таким системам, четыре подхода к определению языка аудиосообщения – акустический, фонотактический, лексический и просодический. Проведен анализ и классификация исследований определения языка аудиосообщения, определены направления развития современных алгоритмов, в том числе наиболее точные и перспективные с точки зрения применения. В рамках задачи определения языка аудиосообщения по аналогии с задачами распознавания диктора можно выделить два направления: идентификацию и верификацию языка аудиосообщения.

На практике чаще решается закрытая задача - текстонезависимая идентификация языка аудиосообщения на замкнутом множестве.

Подходы к определению языка аудиосообщения

Основной проблемой построения САОЯ является поиск решений по построению системы, способной идентифицировать язык на достаточно коротких аудиосообщениях.

На физическом уровне из речи человека можно выделить акустические признаки, такие как частота основного тона, средние значения мощности в спектральных полосах, формантные и другие параметры речевого сигнала.

На элементарном уровне языки различаются набором фонем, акустической реализацией фонем, фонотактике, частоте встречаемости фонем и их комбинаций.

На лексическом уровне языки различаются лексиконами.

На просодическом уровне отличительными чертами являются интонация, тон, тембр и т.д.

Таким образом, опираясь на уровни абстракции речевого сигнала можно выделить четыре подхода - акустический, фонотактический, лексический (распознаватель непрерывной речи с большим словарем – LVCSR), просодический.

Акустический подход основывается на выявлении отличий между языками с помощью моделирования распределений спектральных параметров.

Применяемые параметры: логарифмически масштабированные кепстральные коэффициенты - MFCC, кепстральные коэффициенты линейного предсказания - LPCC, коэффициенты перцептивного линейного предсказания - PLP, смещенные кепстральные коэффициенты – SDC.

Применяемые математические модели: векторное квантование, смеси гауссовых распределений скрытые марковские модели модели опорных векторов модели на основе искусственных нейронных сетей.

В фонотактическом подходе для определения языка используется правила, ограничивающие сочетаемость фонем в различных позициях в составе слова или морфемы, называемые фонотактикой. Для каждого языка характерна своя фонотактика.

Лексический подход. Эффективным является способ идентификации, основанный на распознавателе непрерывной речи с большим словарем. При этом подходе используются целостные данные о лексической и грамматической структуре языка.

Основные сложности состоят в том, что обучение системы должно происходить на правильно транскрибированных данных целевых языков, что не всегда возможно, также требуются большие вычислительные ресурсы системы.

Просодический подход определения языка аудиосообщения основан на использовании просодической или супрасегментной информации речи человека.

На уровне фонетических слов супрасегментным средством является словесное ударение, на уровне фонетической синтагмы – ударение синтагмы, на уровне фразы – интонация, фразовое ударение. Интонация служит для скрепления фразы фонетически, указывает на коммуникативную цель. Фразовое ударение можно представить комбинацией ударений фонетических слов.

В тональных языках в пределах фонетического слова реализуются слоговые тоны, которые отвечают за слоговую просодию.

На практике супрасегментные звуковые средства имеют свою акустическую базу – частотой основного тона речи, длительностью звуков по времени, интенсивностью звуков.

Просодические характеристики обладают свойством устойчивости к изменению акустической обстановки, кратковременной вариативности (неодинаковость произношения ключевой фразы при каждом доступе в систему) и долговременной вариативности (анатомическими изменениями речевого тракта в течение жизни) параметров речеобразующего тракта диктора.

В результате проведенного исследования сделаны выводы о том, что:

- фонотактические системы САОЯ, акустические системы на i-векторах, а также их комбинации (так называемые фьюжн-системы) обеспечивают наилучшие результаты по определению языка аудиосообщения;
- САОЯ, построенные на основе акустического, фонотактического, лексического подходов, а также системы, реализованные их комбинацией, требуют восстановления исходной формы речевого сигнала;
- просодический подход имеет потенциал развития;
- перспективным направлением построения САОЯ по речи, подвергнутой вокодерному преобразованию, без восстановления исходной формы речевого сигнала, является просодический подход.

Поставлена задача научно-обоснованной разработки просодической системы определения языка аудиосообщения по речи, подвергнутой вокодерному преобразованию, без восстановления исходной формы речевого сигнала. В случае большого числа языков предложена реализация смешанной двухэтапной системы, в которой на втором этапе определение языка происходит с восстановлением исходной формы речевого сигнала на основе акустического, фонотактического или лексического подходов.

Кроме того, необходимо решить следующие задачи:

1. анализ работы вокодеров речи;
2. анализ различий между языками на просодическом уровне;
3. разработка способов описания просодических признаков речи человека;
4. выбор математического аппарата для классификации языков по просодическим признакам.

Во второй главе рассматриваются следующие вопросы: анализ работы вокодеров речи, отличия 10 экспериментальных языков на просодическом уровне, использование широких фонетических категорий для описания просодических признаков речи человека, использование акустического подхода для реализации второго этапа определения языка аудиосообщения.

Анализ вокодерного преобразования речи

Анализ алгоритмов работы вокодеров показывает, что в вокодерах передаются такие параметры, как сигнал тон-шум, значения частоты основного тона, усиление. Данные параметры являются акустическими носителями просодических признаков речи и позволяют построить просодическую систему определения языка без восстановления исходной формы речевого сигнала.

Особенности языков на просодическом уровне

Анализ особенностей языков на просодическом уровне позволяет сделать вывод, что различия языков на просодическом уровне заключаются в интонации языка, то есть изменении ЧОТ на ритмической группе, синтагме, фонетическом предложении и фразе, в размерах слов, в месте ударений в словах, наличиях главного и побочного ударений, а также в градациях тональности для тоновых языков. Просодические признаки сложно математически интерпретировать. Для их математической интерпретации предлагается использовать широкие фонетические категории.

Обоснование использования широких фонетических категорий

Вокализованные, невокализованные звуки и паузы как широкие фонетические категории (ШФК – рисунок 1) нашли свое применение в задачах верификации диктора.

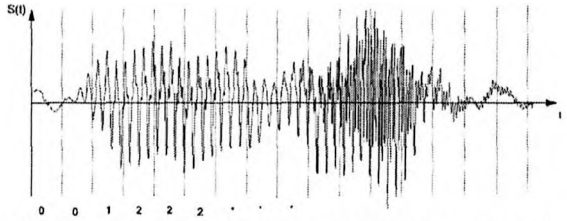

Рисунок 1 – Представление речевого сигнала последовательностью широких фонетических категорий

Традиционно применение ШФК заключается в использовании трех категорий, когда речевой сигнал представляется в виде последовательности квазистационарных сегментов, каждый из которых классифицируется как пауза, вокализованный или невокализованный, и сегменту присваивается номер (соответственно 0, 1 или 2). После получения последовательности чисел, представляющих собой ШФК, от этой последовательности вычисляется автокорреляционная функция.

Описание просодических признаков данной последовательностью имеет ряд недостатков, которые связаны с малой информативностью классификации сегментов только на 3 класса. При данном подходе невозможно выделить возрастание/убывание ЧОТ, главные и побочные максимумы ЧОТ на отрезках различной длительности, взаимозависимость изменения ЧОТ и мгновенной энергии. Это подтвердили исследования, которые показали, что использования 3 категорий недостаточно для надежного определения языка аудиосообщения.

В диссертации предложено расширить множество широких фонетических категорий на основе более детального описания изменения частоты основного тона и кратковременной энергии речевого сигнала.

В третьей главе дается описание разработанных алгоритмов интерпретации просодических признаков – алгоритма на основе широких фонетических категорий и алгоритма на основе кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий, приводится разработанная методика использования данных алгоритмов.

Алгоритм на основе широких фонетических категорий

Количество широких фонетических категорий предлагается расширить и для определения языка аудиосообщения использовать следующий алгоритм.

Пусть множество $L = \{L_1, L_2, ..., L_N\}$ есть множество языков, на котором осуществляется процедура определения языка аудиосообщения, где N – общее число языков. Пусть каждый язык L_i представляется множеством аудиосообщений различных дикторов этого языка $L_i = \{l_1, l_2, ..., l_{Mi}\}$, где Mi – общее число аудиосообщений языка Li.

Аудиосообщение разбивается на квазистационарные сегменты $s_i(m)$ длительностью K отсчетов, где i – i-й сегмент речевого сигнала, i=1,2...P, P – общее число сегментов в аудиосообщении речевого сигнала, $m = 1,..., K-1$. На каждом сегменте i вычисляется признак в соответствии с природой сегмента – вокализованный, невокализованный или пауза $A_i = \mathrm{T}(s_i(m))$, где T – операция вычисления типа сегмента, а также кратковременная энергия сегмента $Ek_i = \mathrm{E}(s_i(m))$, где E – операция вычисления кратковременной энергии сегмента. При работе алгоритма без восстановления исходной формы речевого сигнала параметры A_i и Ek_i берутся из кадров вокодерной передачи. Соответственно формируются последовательности $\vec{A} = (A_1, A_2, ..., A_P)$ и $\vec{Ek} = (Ek_1, Ek_2, ..., Ek_P)$. Если сегмент классифицирован как пауза, то $A_i = 0$, если классифицирован как невокализованный, то $A_i = 1$. На каждом вокализованном сегменте вычисляется частота основного тона $F0_i = \mathrm{F}(s_i(m))$, где F – операция вычисления частоты основного тона, и формируется последовательность $\overrightarrow{F0} = (F0_1, F0_2, ..., F0_P)$. При работе алгоритма без восстановления исходной формы речевого сигнала параметр $F0_i$ берется их кадров вокодерной передачи. Диапазон изменения ЧОТ аудиосообщений разбивается на 5 интервалов. Для вокализованных сегментов каждый сегмент обозначается цифрой в соответствии с тем, в какой интервал ЧОТ попадает значение ЧОТ на данном сегменте. $F0u_i = \mathrm{UF}(\overrightarrow{F0})$,

где $F0u_i$ – уровень ЧОТ, UF – операция вычисления диапазона изменения ЧОТ и кодирования каждого сегмента цифровым обозначением, формируется последовательность $\overrightarrow{F0u} = (F0u_1, F0u_2, ...F0u_P)$ – последовательность из значений ЧОТ на сегментах аудиосообщения. Далее вычисляются сегменты возрастания/убывания кратковременной энергии речевого сигнала $Eu_i = \mathrm{UE}(\overrightarrow{Ek})$, кодирующиеся $Eu_i = (+/-)1$ в зависимости от того, возрастает или убывает энергия соответственно, где UE - операция вычисления возрастания/убывания кратковременной энергии речевого сигнала. Формируется последовательность $\overrightarrow{Eu} = (Eu_1, Eu_2, ..., Eu_P)$. Если данный сегмент относится к участку убыванию кратковременной энергии, цифровое значение ЧОТ умножается на (–1).

Для определения побочных и главных ударений определяется главный и побочный максимумы ЧОТ на отрезке между двумя паузами. Если положение максимума ЧОТ и кратковременной энергии совпадают во времени и максимальны на отрезке, то этот сегмент принимается за главный максимум, если максимумы во времени не совпадают, то сегмент принимается за побочный максимум $MAX_i = \Theta(\overrightarrow{F0u}, \overrightarrow{Eu})$, где Θ - операция определения главного и побочного максимумов ЧОТ и кратковременной энергии. Формируется последовательность $\overrightarrow{MAX} = (MAX_1, MAX_2, ..., MAX_P)$

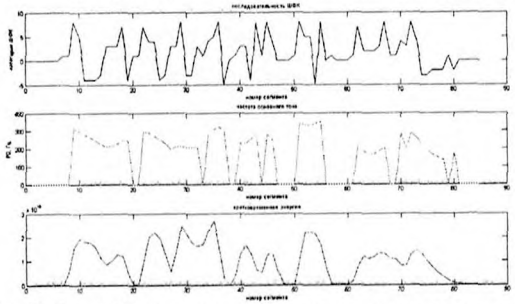

Рисунок 2 – Графики ЧОТ, кратковременной энергии и последовательности ШФК

Таким образом, окончательная последовательность ШФК аудиосообщения $\overrightarrow{X} = (X_1, X_2, ..., X_P)$ состоит из элементов Xi, где

$$X_i = \begin{cases} 0, если & A_i - пауза \\ 1, если & A_i - шум \\ 2, если & F0u_i - уровень\ 1 \\ -2, если & F0u_i - уровень\ 1,\ Eu_i = -1 \\ 3, если & F0u_i - уровень\ 2 \\ -3, если & F0u_i - уровень\ 2,\ Eu_i = -1 \\ 4, если & F0u_i - уровень\ 3 \\ -4, если & F0u_i - уровень\ 3,\ Eu_i = -1 \\ 5, если & F0u_i - уровень\ 4 \\ -5, если & F0u_i - уровень\ 4,\ Eu_i = -1 \\ 6, если & F0u_i - уровень\ 5 \\ -6, если & F0u_i - уровень\ 5,\ Eu_i = -1 \\ 7, если & MAX_i - побочный\ максимум \\ 8, если & MAX_i - главный\ максимум \end{cases}$$

На рисунке 2 представлены 3 графика – частоты основного тона, кратковременной энергии сигнала, последовательности широких фонетических категорий $\vec{X}$. На рисунке 3 и 4 – блок-схема алгоритма кодирования сегментов речевого сигнала.

По последовательности широких фонетических категорий $\vec{X}$ вычисляется автокорреляционная функция $\vec{R} = \Psi(\vec{X})$, где Ψ - операция вычисления автокорреляционной функции. В случае работы алгоритмов без восстановления исходной формы речевого сигнала значения ЧОТ берутся из кадров вокодерной передачи. В случае работы алгоритма с восстановлением исходной формы речевого сигнала требуется выбор алгоритма оценки частоты основного тона.

Для определения ЧОТ существуют различные алгоритмы. В данной диссертационной работе были проведены испытания готовых алгоритмов, реализуюших определение ЧОТ по АКФ – алгоритм SIFT, по КФСР – алгоритм AMDF, а также алгоритм оценки ЧОТ из алгоритма кодирования речи MELP.

Процент отрезков речевого сигнала с показателями Р(ОТ) - правильно определенным ОТ, Р(НВ/В) – принятия вокализованного отрезка за невокализованный, Р(В/НВ) – принятия невокализованного за вокализованный составил:

Алгоритм	SIFT	AMDF	MELP
P (OT), %	87±1	89±1	95±1,5
P (НВ/В), %	7±1	6±1	3±0.5
P (В/НВ), %	0.5	0.5	0.5

Алгоритм на основе кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий

Для реализации просодической классификации предлагается использование кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий сигналов аудиосообщений. В ходе предварительных экспериментов установлено, что предложенный подход обеспечивает повышение достоверности просодической классификации. Аудиосообщение разбивается на квазистационарные сегменты $s_i(m)$ длительностью K отсчетов, где i – i-й сегмент речевого сигнала, i=1,2...P, P – общее число сегментов в аудиосообщении, $m = 1,..., K-1$. На каждом сегменте i вычисляется признак в соответствии с природой сегмента – вокализованный, невокализованный или пауза $A_i = \mathrm{T}(s_i(m))$, i=1,2...P, где T – операция вычисления типа сегмента, а также кратковременная энергия сегмента $Ek_i = \mathrm{E}(s_i(m))$, i=1,2...P, где E – операция вычисления кратковременной энергии сегмента. Соответственно формируются последовательности $\vec{A} = (A_1, A_2, ..., A_P)$ и $\vec{Ek} = (Ek_1, Ek_2, ..., Ek_P)$. При работе алгоритма без восстановления исходной формы речевого сигнала параметры A_i и Ek_i берутся из кадров вокодерной передачи. Если сегмент классифицирован как пауза, то $A_i = 0$, если классифицирован как невокализованный, то $A_i = 1$. На каждом вокализованном сегменте вычисляется частота основного тона $F0_i = \mathrm{F}(s_i(m))$, i=1,2...P, где F – операция вычисления частоты основного тона, и формируется последовательность $\vec{F0} = (F0_1, F0_2, ..., F0_P)$. При работе алгоритма без восстановления исходной формы речевого сигнала параметр $F0_i$ берется их кадров вокодерной передачи.

По последовательности значений частоты основного тона и последовательности кратковременных энергий вычисляется их кросс-корреляционная фуикция $\vec{B} = \Phi(\vec{F0}, \vec{Ek})$, где Φ - операция вычисления кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий. Вектор значений кросскорреляционной функции последовательности широких фонетических категорий подается на вход нейронной сети, которая принимается решение по отнесению данного вектора к какой-либо группе языков.

Алгоритм вычисления признаков представлен на рисунке 5.

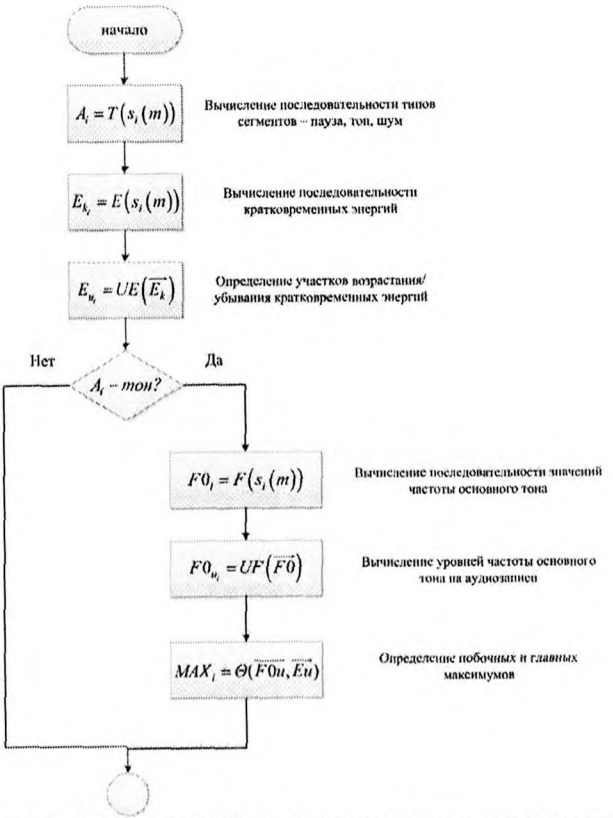

Рисунок 3 – Блок-схема алгоритма кодирования сегментов речевого сигнала

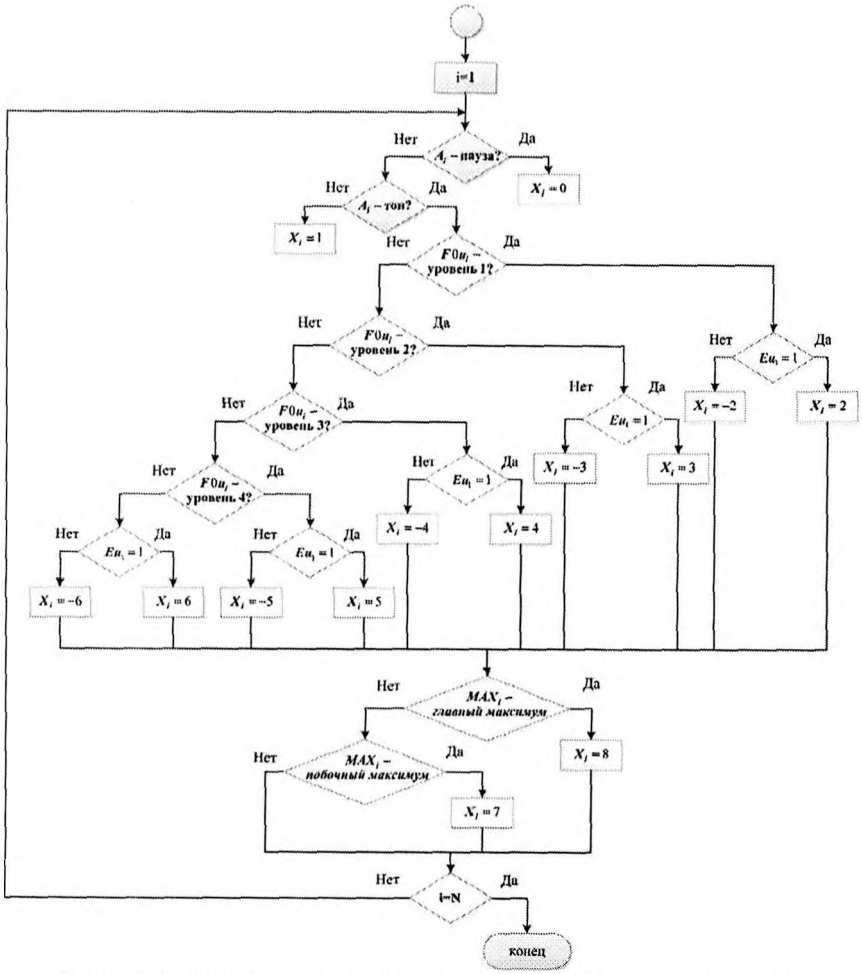

Рисунок 4 – Блок-схема (продолжение) алгоритма кодирования сегментов речевого сигнала

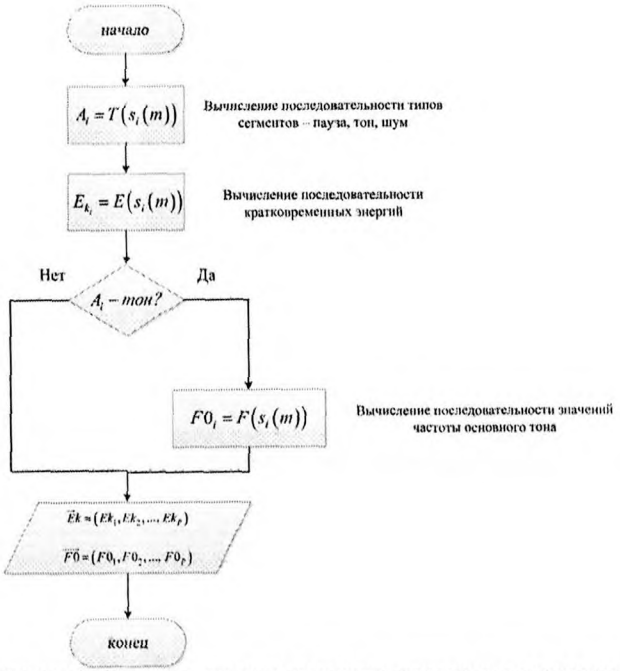

Рисунок 5 – Блок-схема алгоритма кодирования сегментов речевого сигнала

Методика использования алгоритмов интерпретации просодических признаков речи

Для применения указанных алгоритмов была разработана следующая методика. Она заключается в последовательности ряда этапов.

Этап 1. Формирование обучающей речевой базы данных. Обучающая база данных должна удовлетворять следующим условиям: если N – общее число языков, d_m^i – число дикторов мужского пола языка i, d_f^i – число дикторов женского пола языка i, то $V_i(d_m^i, d_f^i) = V_j(d_m^j, d_f^j)$, где i, j = номера языков, i,j=1…N. То есть все возрастные группы должны быть представлены в равной пропорции дикторами мужского и женского пола, объемы речевых данных дикторов различных возрастных групп должны быть одинаковы. Объем речевых данных должен быть достаточен со статистической точки зрения для описания всех вариативностей произношения на данном языке. Общие объемы речевых баз по языкам должны быть равны.

Шаг 1. Получение от источника аудиосообщения в цифровом виде $S_i(f_d,m,p,f_r)$ с параметрами – формат f_r = «wav», частота дискретизации f_d = 8 кГц, режим m = моно, p = 16 бит, t-номер аудиосообщения.

Шаг 2. Фильтрация аудиосообщения $S_i(f_d,m,p,f_r)$ – удаление посторонних шумов. Получение фильтрованного аудиосообщения $S_i^f(f_d,m,p,f_r)=P[S_i(f_d,m,p,f_r)]$, где P – операция фильтрации.

Шаг 3. Формирование обучающих и тестовых данных. Для каждого языка L_i формируется база аудиосообщений $Z_{Li}\{S_1^f{}_{Li}(f_d,m,p,f_r), S_2^f{}_{Li}(f_d,m,p,f_r),\ldots, S_{Mi}^f{}_{Li}(f_d,m,p,f_r)\}$, где Mi – общее число аудиосообщений языка Li.

Общая база аудиосообщений $Z=\{ Z_{L1}, Z_{L2},\ldots, Z_{LN} \}$.

Шаг 4. Обработка всех аудиосообщений всех языков заданным вокодером.

$Z^{vok}=VOK(Z)$, где VOK – операция обработки базы аудиосообщений вокодером. $Z^{vok}=\{ Z^{vok}_{L1}, Z^{vok}_{L2},\ldots, Z^{vok}_{LN} \}$.

Шаг 5. Вычисление параметров из аудиосообщений в соответствии с разработанными алгоритмами – формирование базы параметров $Z^{vok}_{Li}{}^{Mod1}=Mod1(Z^{vok}_{Li})$, $Z^{vok}_{Li}{}^{Mod2}=Mod2(Z^{vok}_{Li})$, где Mod1, Mod2 – операции вычисления параметров в соответствии с разработанными алгоритмами описания просодических параметров речи.

Этап 2. Обучение искусственной нейронной сети, в процессе обучения происходит настройка различных параметров нейронной сети. Нейронные сети с различной топологией описываются различными математическими моделями, поэтому в каждом конкретном случае нейронная сеть будет описываться своей формулой. Нейронные сети строятся попарно для каждой пары языков Li и Lj, i≠j.

Этап 3.Тестовая оценка нейронной сети

Шаг 1. Получение от источника аудиосообщения в цифровом виде $S(f_d,m,p,f_r)$ с параметрами – формат f_r = «wav», частота дискретизации f_d = 8 кГц, режим m = моно, p = 16 бит.

Шаг 2. Фильтрация аудиосообщения $S(f_d,m,p,f_r)$ – удаление посторонних шумов. Получение фильтрованного аудиосообщения $S^f(f_d,m,p,f_r)=P[S(f_d,m,p,f_r)]$, где P – операция фильтрации.

Шаг 3. Обработка аудиосообщения заданным вокодером $S^{vok}=VOK(S^f(f_d,m,p,f_r))$, где VOK – операция обработки аудиосообщения вокодером.

Шаг 4. Вычисление параметров из S^{vok} в соответствии с разработанными алгоритмами $S^{vok}{}_{Mod1}=Mod1(S^{vok})$, $S^{vok}{}_{Mod2}=Mod2(S^{vok})$, где Mod1, Mod2 –операции вычисления параметров в соответствии с разработанными алгоритмами описания просодических параметров речи.

Шаг 5. Тестирование нейронных сетей.

$$\hat{L}_t = NET(S^{vok\,Mod1}),$$

где NET – операция тестирования нейронной сети, на выходе которой $\hat{L}_t$ - оценка языка.

Шаг 6. При построении различных архитектур многослойного персептрона для пар языков формируется вектор целевых показателей достоверности распознавания $D = (d_{1,2}, d_{2,1}, d_{1,3}, d_{3,1}, d_{i,j}, d_{j,i}, ..., d_{N,N-1}, d_{N-1,N})$, где N – общее число языков в САОЯ. Каждый элемент $d_{i,j}$, $d_{j,i}$ = 100.

Вычисляется число правильно распознанных аудиосообщений в каждой паре языков. Получается вектор

$D_k = (d^k_{1,2}, d^k_{2,1}, d^k_{1,3}, d^k_{3,1}, d^k_{i,j}, d^k_{j,i}, ..., d^k_{N,N-1}, d^k_{N-1,N})$, где d^k_{ij}-число правильно определенных аудиосообщений для пары языков L_iL_j, i≠j.

Расстояние между D и D_k определяется как

$$D_r = \sqrt{\left(d_{1,2}-d^k_{1,2}\right)^2+\left(d_{2,1}-d^k_{2,1}\right)^2+\left(d_{i,j}-d^k_{i,j}\right)^2+\left(d_{j,i}-d^k_{j,i}\right)^2 ...+\left(d_{N,N-1}-d^k_{N,N-1}\right)^2+\left(d_{N-1,N}-d^k_{N-1,N}\right)^2}$$

Таким образом, тем меньше расстояние D_r, тем лучше настроена НС.

Шаг 7. Построение иерархического дерева языков на основе агломеративного иерархического алгоритма

$$\rho_{min}(\omega_i,\omega_j)=\min_{x_k\in\omega_i, x_l\in\omega_j} d(x_k,x_l),$$

где ω_i,ω_j - языки Li и Lj, $\rho(\omega_i,\omega_j)$ - расстояние между Li и Lj.

На основе иерархического дерева строятся группы языков в случае двухэтапного алгоритма.

В четвертой главе проводится экспериментальная оценка разработанных моделей, для этого формируется база аудиосообщений на целевых языках, выбирается и реализуется решающее правило.

Формирование речевой базы данных

Для проведения экспериментальной оценки моделей была сформирована база данных аудиосообщений, состав базы указан в таблице № 1.

Источник аудиосообщений – каналы интернет вещания – телевидение и радио, то есть речь, прошедшая обработку различными кодеками.

Для исключения влияния базы данных на эксперимент число дикторов по всем языкам выбрано одинаковым, суммарное время аудиосообщений выбрано одинаковым, также одинаков процент обучающей и тестовой выборок. Обучающая и тестовая выборки не перекрываются. Для проведения экспериментов все аудиосообщения обучающей и тестовой выборок разделялись на отрезки по 10 секунд. Выбор аудиозаписей для обучения и тестирования производился в случайном порядке.

Возрастной состав дикторов определить возможно приблизительно – мужчины и женщины от 20 до 50 лет.

Таблица 1 – Характеристики базы данных для проведения экспериментальной оценки эффективности алгоритмов интерпретации просодических признаков

Язык	Число дикторов	Суммарное время аудиозаписей на каждого диктора, мин	Пол диктора (м-мужской, ж-женский)	Процент обучающей/тестовой выборки, %
Китайский	10	100	5м/5ж	80/20
Английский	10	100	5м/5ж	80/20
Финский	10	100	5м/5ж	80/20
Французский	10	100	5м/5ж	80/20
Немецкий	10	100	5м/5ж	80/20
Японский	10	100	5м/5ж	80/20
Персидский	10	100	5м/5ж	80/20
Португальский	10	100	5м/5ж	80/20
Русский	10	100	5м/5ж	80/20
Испанский	10	100	5м/5ж	80/20

Создание и настройка нейронной сети

В данной работе для классификации отрезков речи применены искусственные нейронные сети.

Как известно, для задач типа классификации число нейронов во входном слое вычисляется исходя из вектора признаков, который подается на вход, а число нейронов выходного слоя зависит от того, какая задача решается и какое применяется правило интерпретации выходных значений. Для оценки числа нейронов в скрытых слоях применялась формула

$$\frac{N_y N_p}{1+log_2(N_p)} \le N_w \le N_y\left(\frac{N_p}{N_x}+1\right)(N_x+N_y+1)+N_y$$

где Ny – размерность выходного вектора НС, Np – число элементов обучающей выборки, Nx – размерность входного вектора, Nw – общее число нейронов.

Рассмотрим обучение НС. Если X – входной сигнал НС, Y – выходной сигнал НС, то НС реализует функцию G:

$$Y = F(X)$$

Функция G определяется архитектурой НС, смещениями в НС и синаптическими весами. Если заданы входные-выходные пары данных (X_1,Y_1), (X_2,Y_2), … (X_N,Y_N), при этом $Y_i=F^*(X_i)$, то обучение НС заключается в поиске функции F, совпадающей с F* с точностью до ошибки E.

В данной работе применяется «обучение с учителем». Поскольку для обучения сети используется лишь набор примеров, то поэтому и вся информация заключается в данном наборе примеров, а следовательно и от данного набора зависит качество обучения.

Переобучение НС заключается в том, что при наличии большого числа весов НС моделирует все более сложную функцию, при этом может иметь значения в тех точках, в которых он существовать не должен. В связи с этим для исключения переобучения используется механизм перекрестной проверки, при котором часть данных в обучении не участвует, а используется для независимой проверки результата. В некоторых случаях, и не только при обучении НС, используются 3 множества – обучающее, подтверждающее и тестовое с примерным соотношением обучающих, подтверждающих и тестовых данных 60/20/20 % либо 70/20/10 %.

В связи с этим, как показывает практика и отсутствие точных методов выбора архитектуры и весов сетей, для решения конкретной задачи необходимо проводить экспериментальное исследование различных сетей по типу и архитектуре и проводя настройку столько раз, сколько потребуется для достижения малой ошибки.

Для реализации классификатора на базе нейронной сети был сделан выбор в пользу пакета MATLAB, который включает в себя функционал по нейронным сетям.

В диссертации экспериментальные исследования проводились со следующими сетями: сеть кохонена, каскадная нс, сеть элмана, многослойный персептрон, сеть хопфилда, вероятностная сеть, сеть с радиальными базисными функциями RBF, нейросети встречного распространения – LVQ сети.

Алгоритмы, стандартные в MATLAB, использованные при обучении сетей: квазиньютоновский алгоритм; алгоритм Левенберга-Марквардта с регуляризацией Байеса; метод сопряженных градиентов Флетчера-Ривса; метод сопряженных градиентов Полака-Ривьера; метод сопряженных градиентов Пауэлла-Беаля; базовый метод градиентного спуска; метод градиентного спуска с переменным шагом обучения; алгоритм Левенберга-Марквартда, метод масштабированных сопряженных градиентов; метод градиентного спуска с моментом; метод градиентного спуска с моментом и переменным шагом обучения; метод «One Step Secant»; метод случайных приращений; эластичный алгоритм обратного распространения ошибки.

В результате проведения предварительных вычислительных экспериментов с целью выбора типа НС с указанными сетями для оценки двух указанных ранее моделей, все НС показали неудовлетворительные результаты определения языка. Наилучшие показатели были получены при создании многослойного персептрона, поэтому было принято решение провести более точную настройку данного типа НС.

Следует отметить, что есть различные варианты построения НС – либо одна НС для всех языков, либо несколько НС для различных групп. Но поскольку заранее неизвестно, какой язык подается на вход НС, было принято решение строить одну общую НС для всех пар языков.

Оценка модели на основе широких фонетических категорий

Согласно формуле и исходным параметрам для тестирования НС: N_y=2, N_p=600, N_x=399, число нейронов в скрытых слоях $117 \le N_w \le 2015$ для модели ШФК.

Поскольку N_w лежит в пределах от 117 до 2015, то при создании архитектуры НС число нейронов в слое варьировалось от 100 до 2000, соответственно число слоев от 1 (1 слой от 100 до 2000 нейронов) до 20 (20 слоев по 100 нейронов) в следующих конфигурациях: со 100 до 1000 нейронов с шагом в 10 нейронов в слое, с 1000 до 2000 с шагом в 100 нейронов. Максимальное число нейронов в слое 800

При построении различных архитектур многослойного персептрона для 45 пар языков формировался вектор целевых показателей достоверности распознавания $D = (d_{1,2}, d_{2,1}, d_{1,3}, d_{3,1}, d_{i,j}, d_{j,i}, \dots, d_{N,N-1}, d_{N-1,N})$, где N – общее число языков в САОЯ. Таким образом, длина вектора D = 90. Каждый элемент $d_{i,j}$, $d_{j,i}$ = 100.

Вектор показателей достоверности распознавания $D_k = (d^k{}_{1,2}, d^k{}_{2,1}, d^k{}_{1,3}, d^k{}_{3,1}, d^k{}_{i,j}, d^k{}_{j,i}, \dots, d^k{}_{N,N-1}, d^k{}_{N-1,N})$ для текущей архитектуры НС имеет также длину 90 и расстояние между D и D_k определяется как

$$D_r = \sqrt{\left(d_{1,2} - d^k{}_{1,2}\right)^2 + \left(d_{2,1} - d^k{}_{2,1}\right)^2 + \left(d_{i,j} - d^k{}_{i,j}\right)^2 + \left(d_{j,i} - d^k{}_{j,i}\right)^2 \dots + \left(d_{N,N-1} - d^k{}_{N,N-1}\right)^2 + \left(d_{N-1,N} - d^k{}_{N-1,N}\right)^2}$$

Таким образом, тем меньше расстояние D_r, тем лучше настроена НС. В результате исследования Dr колебалась в пределах от 59.1861 до 532.4106. Наилучший показатель Dr=72.5358 был получен для конфигурации НС – общее число нейронов 1400, 1 слой 800 нейронов, 2 слой 600 нейронов.

Результаты определения языка представлены в таблице 2.

Таблица 2 - Средние значения достоверности определения языка

	китайский	английский
китайский	94,5±1,8	
английский		93,8±1,5

	китайский	финский
китайский	95,1±1,6	
финский		93,8±1,5

	китайский	французский
китайский	96,2±0,9	
французский		94,2±1,1

	китайский	немецкий
китайский	95,9±1,4	
немецкий		94,5±0,6

	китайский	японский
китайский	97,5±1,5	
японский		83,6±1,1

	китайский	персидский
китайский	96,6±1,2	
персидский		84,4±0,8

	китайский	португальский
китайский	95,2±1,3	
португальский		94,2±1,3

	китайский	русский
китайский	94,4±1,6	
русский		94,4±1,8

	китайский	испанский
китайский	97,9±0,7	
испанский		93,9±1,7

	английский	финский
английский	97,4±1,4	
финский		93,7±1,1

	английский	французский
английский	92,8±1	
французский		93,6±1,1

	английский	немецкий
английский	93,8±0,8	
немецкий		92,6±1,7

	английский	японский
английский	93,6±1,2	
японский		94,1±1

	английский	персидский
английский	98,1±1,5	
персидский		94±1,3

	английский	португальский
английский	94,5±1,7	
португальский		93,6±1,4

	английский	русский
английский	94±1,1	
русский		95,1±0,8

	английский	испанский
английский	97,8±1,5	
испанский		94,3±1,5

	финский	французский
финский	93,2±1,4	
французский		93,2±1,4

	финский	немецкий
финский	93,4±1,6	
немецкий		93,7±1,2

	финский	японский
финский	93,9±1,4	
японский		74±0,9

	финский	персидский
финский	93,9±1,5	
персидский		74,6±1,4

	финский	португальский
финский	96,1±1,4	
португальский		93,5±1

	финский	русский
финский	93,7±1	
русский		94,1±1,5

	финский	испанский
финский	94,3±1	
испанский		93,4±1,4

	французский	немецкий

	французский	японский

	французский	персидский

французский	93,9±0,8	
немецкий		92,5±1,7

французский	93,4±1,3	
японский		98,3±1,4

французский	94±1,5	
персидский		93,3±1,5

	французский	португальский
французский	94,8±0,8	
португальский		93,9±2,2

	французский	русский
французский	93,8±1,4	
русский		95,3±2

	французский	испанский
французский	94,4±1,1	
испанский		93,2±1,6

	немецкий	японский
немецкий	94,6±1,3	
японский		93,3±1,8

	немецкий	персидский
немецкий	94±1,3	
персидский		93,8±1,1

	немецкий	португальский
немецкий	97,5±1	
португальский		94,2±0,8

	немецкий	русский
немецкий	96,3±1,4	
русский		93,4±1,3

	немецкий	испанский
немецкий	93,9±1,4	
испанский		94,2±1,2

	японский	персидский
японский	94±1,4	
персидский		83,6±1,1

	японский	португальский
японский	84,9±1,5	
португальский		94,5±1,6

	японский	русский
японский	94,4±1,5	
русский		94±1,2

	японский	испанский
японский	98±1,1	
испанский		93,8±1,6

	персидский	португальский
персидский	92,7±1,5	
португальский		93,5±1,3

	персидский	русский
персидский	84,3±1,1	
русский		94,4±1,1

	персидский	испанский
персидский	93,2±1,6	
испанский		94,1±1,3

	португальский	русский
португальский	93,9±1	
русский		94,3±2

	португальский	испанский
португальский	98,4±1,6	
испанский		94,5±0,9

	русский	испанский
русский	94,5±1,6	
испанский		93,2±1,4

Оценка модели кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий

Согласно формуле и исходным параметрам для тестирования НС: N_y=2, N_p=600, N_x=797, число нейронов в скрытых слоях $117 \le N_w \le 2806$ для модели ВКФ от ОТ и КЭ.

Поскольку N_w лежит в пределах от 117 до 2806, то при создании архитектуры НС число нейронов в слое варьировалось от 100 до 3000, соответственно число слоев от 1 (1 слой от 100 до 3000 нейронов) до 20 (30 слоев по 100 нейронов) в следующих конфигурациях: со 100 до 1000 нейронов с шагом в 10 нейронов в слое, с 1000 до 3000 с шагом в 100 нейронов. Максимальное число нейронов в слое 800. Dr=89.1449.

Результаты определения языка представлены в таблице 3.

Таблица 3 - Средние значения достоверности определения языка

	китайский	английский
китайский	97,7±1,5	
английский		91,2±1,9

	китайский	финский
китайский	94,7±1,8	
финский		90,9±1,9

	китайский	французский
китайский	92,8±1,9	
французский		92,1±1,5

	китайский	немецкий
китайский	97,8±1,6	
немецкий		92,5±2,5

	китайский	японский
китайский	97,9±1,5	
японский		80,6±1,8

	китайский	персидский
китайский	93,8±2	
персидский		71,1±2,6

	китайский	португальский
китайский	91,7±1,8	
португальский		90,7±2,8

	китайский	русский
китайский	93,1±1,7	
русский		91±2,1

	китайский	испанский
китайский	92,1±1,3	
испанский		90,5±2

	английский	финский
английский	91,4±1,9	
финский		91,5±2

	английский	французский
английский	92,3±1,8	
французский		92,9±1,7

	английский	немецкий
английский	92,9±1,6	
немецкий		90,2±1,8

	английский	японский
английский	94,8±1,9	
японский		91,8±1,6

	английский	персидский
английский	92,7±2,5	
персидский		91,5±1,7

	английский	португальский
английский	97,7±1,2	
португальский		91±1,9

	английский	русский
английский	90,3±1,4	
русский		91,7±2,1

	английский	испанский
английский	92±1,5	
испанский		92,9±1,5

	финский	французский
финский	95,8±1,8	
французский		92,4±1,8

	финский	немецкий
финский	94,7±1,1	
немецкий		91,4±1,6

	финский	японский
финский	94,6±2,3	
японский		90,1±1,8

	финский	персидский
финский	95,4±2,5	
персидский		82,3±1

	финский	португальский
финский	90,9±2	
португальский		92±2,1

	финский	русский
финский	93,6±2,4	
русский		90,6±2,4

	финский	испанский
финский	95,9±1,5	
испанский		90,9±1,8

	французский	немецкий
французский	93,9±1,1	
немецкий		90,4±1,1

	французский	японский
французский	96,7±1,8	
японский		82,3±1,1

	французский	персидский
французский	97,5±1,8	
персидский		91,6±2

	французский	португальский
французский	92,1±1,5	
португальский		92±1

	французский	русский
французский	91,8±1,8	
русский		92,6±1,4

	французский	испанский
французский	91,8±1,4	
испанский		92,8±1,9

	немецкий	японский
немецкий	91,8±1,1	
японский		71,9±2,8

	немецкий	персидский
немецкий	92,2±1,6	
персидский		82,6±1,3

	немецкий	португальский
немецкий	92,4±2,8	
португальский		93,2±1,1

	немецкий	русский

	немецкий	испанский

	японский	персидский

немецкий	93±2,4	
русский		92,3±1,8

немецкий	95,4±3,1	
испанский		91,2±1,8

японский	90,7±1,6	
персидский		78,2±1,8

	японский	португальский
японский	90,5±1,8	
португальский		93,4±1,1

	японский	русский
японский	94,7±2,1	
русский		92,4±1,3

	японский	испанский
японский	97,2±1,2	
испанский		93,1±1,4

	персидский	португальский
персидский	97,5±1,8	
португальский		92,1±1,2

	персидский	русский
персидский	92,5±2	
русский		91,7±2

	персидский	испанский
персидский	91,5±2,2	
испанский		91,4±1,4

	португальский	русский
португальский	94,5±2,4	
русский		91,6±2,7

	португальский	испанский
португальский	92,5±1,7	
испанский		92,1±1,9

	русский	испанский
русский	96,2±2,2	
испанский		93,6±2

В заключении приводятся результаты диссертационного исследования, а также перспективы дальнейших исследований.

В ходе проведенных исследований решены следующие задачи и сделаны выводы:

1. Проведена классификация систем определения языка аудиосообщения, анализ существующих подходов определения языка аудиосообщения – акустического, фонотактического, лексического и просодического. В результате анализа установлено, что фонотактические системы автоматического определения языка аудиосообщения, акустические системы на i-векторах, а также их комбинации (так называемые фьюжн-системы) обеспечивают наилучшие результаты по определению языка аудиосообщения, просодический подход развивается, он ортогонален с точки зрения используемых параметров акустическому, фонотактическому и лексическому подходам. Системы определения языка, построенные на основе акустического, фонотактического, лексического подходов, а также системы, реализованные их комбинацией, требуют восстановления исходной формы речевого сигнала;
2. Анализ исследований систем автоматического определения языка показал, что перспективным направлением построения систем автоматического определения языка по речи, подвергнутой вокодерному преобразованию, без восстановления исходной формы речевого сигнала, является просодический подход.
3. Проведен анализ работы алгоритмов работы вокодеров. В результате анализа установлено, что в вокодерах передаются такие параметры, как сигнал тон-шум, значения частоты основного тона, усиление. Данные параметры являются акустическими носителями просодических признаков речи и позволяют построить просодическую систему определения языка без восстановления исходной формы речевого сигнала.
4. Проведено исследование отличий языков на просодическом уровне. Установлено, что различия языков на просодическом уровне заключаются в интонации языка, то есть изменении частоты основного тона на ритмической группе, синтагме, фонетическом предложении и фразе, в размерах слов, в месте ударений в словах, наличиях главного и побочного ударений, а также в градациях тональности для

тоновых языков. Для комплексного описания просодических признаков предложено использовать широкие фонетические категории.

5. Проведено исследование информативности трех широких фонетических категорий и мелодики основного тона для описания просодических признаков речи диктора. В результате исследования сделан вывод, что для комплексного описания просодических признаков необходимо увеличение числа ШФК.

6. Разработан алгоритм интерпретации просодических признаков речи диктора в виде автокорреляционной функции от последовательности широких фонетических категорий, не требующий восстановления исходной формы речевого сигнала.

7. Разработан алгоритм интерпретации просодических характеристик речи диктора в виде кросскорреляционной функции от последовательности частот основного тона и последовательности кратковременных энергий, не требующий восстановления исходной формы речевого сигнала.

8. Разработана методика использования алгоритмов интерпретации просодических признаков речи в задаче определения языка аудиосообщения по речи, преобразованной вокодерами, без восстановления исходной формы речевого сигнала.

9. Проведена экспериментальная оценка надежности определения языка аудиосообщения на основе разработанных алгоритмов и методики. Проведена оценка репрезентативности полученных результатов.

Основные новые научные и практические результаты:

1. Алгоритм интерпретации просодических характеристик речи диктора в виде автокорреляционной функции от последовательности широких фонетических категорий, отличающийся от известных алгоритмов тем, что не требует восстановления исходной формы речевого сигнала.

2. Алгоритм интерпретации просодических характеристик речи диктора в виде кросскорреляционной функции от последовательности частоты основного тона и кратковременной энергии, отличающийся от известных алгоритмов тем, что не требует восстановления исходной формы речевого сигнала.

3. Методика использования алгоритмов интерпретации просодических признаков в задачах определения языка аудиосообщения по речи, преобразованной вокодерами, без восстановления исходной формы речевого сигнала.

4. Оценена эффективность разработанных алгоритмов, проведена оценка репрезентативности результатов.

Разработанные алгоритмы отличаются от известных алгоритмов тем, что:

1. комплексно описывают просодические признаки речи, что позволяет использовать данные признаки при специальной обработке аудиосообщений;

2. не требуют восстановления исходной формы речевого сигнала, поэтому применимы при передаче речи по узкополосным каналам связи с помощью низкоскоростных вокодеров, поскольку используемые акустические параметры передаются в низкоскоростных вокодерах без изменений.

В **приложении 1** приведены графики изменения ЧОТ во времени для различных языков. В **приложении 2** приведен фрагмент кода программы алгоритмов интерпретации просодических признаков, в **приложении 3** приведены акты о внедрении результатов диссертационного исследования.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в журналах, входящих в перечень рецензируемых научных изданий, и рекомендованных Высшей аттестационной комиссией при Министерстве образования и науки Российской Федерации для опубликования основных научных результатов диссертаций:

1. Бессонов, М.А. Шалимов, И.А. Обзор методов автоматической идентификации языка аудиосообщения // Труды НИИР: сб. ст. – М.: НИИР. 2011. № 3. – с. 43-47

2. Бессонов, М.А., Шалимов, И.А. Анализ состояния и перспектив развития технологий определения языка аудиосообщения // Труды НИИР: сб.ст. – М.: НИИР, 2013. № 3, стр. 23-30

3. Бессонов, М.А., Шалимов, И.А. Сомов, А.М. Определение языка аудиосообщения на основе акустических параметров речи диктора. // Труды НИИР: сб. ст. – М.: НИИР. 2013. № 3. – с.18-22.

4. Бессонов, М.А., Шалимов, И.А., Костенко, А.И., Босомыкин, Д.В. О повышении достоверности определения языка аудиосообщения на основе просодической классификации // Труды НИИР: сб.ст. – М.: НИИР, 2014. № 4, стр. 2-7

5. Бессонов, М.А., Фархадов, М.П. Алгоритмы интерпретации просодических признаков речи при ее обработке низкоскоростными кодеками / Управление большими системами. Выпуск 66. М.: ИПУ РАН, 2017. С.6-24.

Публикации в материалах научных мероприятий

6. Бессонов, М.А. Алгоритм описания просодических признаков и его применение // Физика и радиоэлектроника в медицине и экологии. Труды 11-й международной научной конференции «ФРЭМЭ'2014» с элементами научной молодежной школы: Материалы. – Владимир, 2014. с. 63-66

Личный вклад автора в работах, написанных в соавторстве, состоит в следующем. В работе [1] автором проведен анализ существующих подходов определения языка аудиосообщения, в работе [2] автором проведен обзор, анализ существующих и перспективных систем определения языка аудиосообщения, в работе [3] автором предложен вариант акустической системы определения языка, в работе [4, 5, 6] автором предложены: алгоритм интерпретации просодических признаков речи диктора в виде автокорреляционной функции от последовательности широких фонетических категорий, не требующий восстановления исходной формы речевого сигнала; алгоритм интерпретации просодических характеристик речи диктора в виде кросскорреляционной функции от последовательности частот основного тона и последовательности кратковременных энергий, не требующий восстановления исходной формы речевого сигнала; методика использования алгоритмов интерпретации просодических признаков речи в задаче определения языка аудиосообщения по речи, преобразованной вокодерами, без восстановления исходной формы речевого сигнала; проведена экспериментальная оценка надежности определения языка аудиосообщения на основе разработанных алгоритмов и методики; проведена оценка репрезентативности полученных результатов.

Бессонов Максим Александрович

Алгоритмы интерпретации просодических признаков речи при обработке аудиосообщений

В диссертации проведен анализ существующих подходов определения языка аудиосообщения, проанализированы научные публикации по данной тематике, выявлены достоинства и недостатки существующих подходов, в том числе при обработке аудиосообщений без восстановления исходной формы речевого сигнала. Сделан вывод о целесообразности применения просодического подхода, для чего разработаны новые алгоритмы интерпретации просодических признаков речи, разработана методика их применения, проведена экспериментальная оценка алгоритмов. Полученные результаты внедрены в коммерческих компаниях, специализирующихся на обработке речи, а также в ряде ВУЗов в лекционных курсах по обработке речевых сигналов.

Bessonov Maxim Aleksandrovich

Algorithms of interpreting speech prosodic features when processing audio messages

The thesis shows the analysis of the existing systems identifying the language of audio messages, reviews scientific publications on the subject and reveals advantages and disadvantages of existing systems, including the methods of processing audio messages without restoring the original form of speech. The usage of prosodic system is concluded to be reasonable that triggered development of new algorithms for interpreting speech prosodic features, the methods of their application are developed; experimental evaluation of the algorithms is carried out. The obtained results are used in commercial companies, specializing in speech processing as well as in a few Universities in the framework of lectures devoted to speech signals processing.

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Подписано в печать __.___.2017
Тираж ___ экз. Заказ № __.