СОДЕРЖАНИЕ

Список основных сокращений.....3
Введение.....4
ГЛАВА 1. Задача автоматического распознавания речи.....17
1.1. Постановка задачи автоматического распознавания речи. Теоретико-вероятностный подход.....17
1.2. Теоретико-информационный подход. Критерий минимального информационного рассогласования.....22
1.3. Алгоритм распознавания изолированных слов.....27
Выводы по главе.....35
ГЛАВА 2. Разработка адаптивной кластерной модели элементарных речевых единиц.....37
2.1. Задача фонетического анализа речи.....37
2.2. Результаты математического моделирования алгоритма фонетического анализа речи на основе адаптивной кластерной модели.....39
2.3. Анализ эффективности алгоритма.....43
Выводы по главе.....48
ГЛАВА 3. Результаты экспериментальных исследований алгоритма фонетического анализа речи на основе адаптивной кластерной модели…50
3.1. Лабораторный образец фонетического анализатора речи.....51
3.2. Повышение точности алгоритма распознавания речи.....57
Выводы по главе.....61
ГЛАВА 4. Применение адаптивной кластерной модели в новой технологии информационного обеспечения при обучении речи слабослышащих……61
4.1. Обзор существующих подходов к решению задачи обучения речи с визуализацией речевого сигнала.....61
4.2. Визуализация речевого сигнала на основе кластерной модели речевых единиц.....64
Выводы по главе.....87

Заключение.....90
Список литературы.....93

Список основных сокращений

АКМ - автокорреляционная матрица АР — авторегрессия

АРР — автоматическое распознавание речи БД - база данных

ВИР — взаимное информационное рассогласование ДП - динамическое программирование ИНС - искусственная нейронная сеть ИР — информационное рассогласование

ИСОР - информационная система обучения речи глухонемых и слабослышащих

ИС ФАР - информационная система фонетического анализа слитной речи

ИТВР - информационная теория восприятия речи ИТОР - информационная теория обучения речи ИЦ - информационный центр

МИР - минимальное информационное рассогласование

МОФ - метод обеляющего фильтра

ОФ - обеляющий фильтр

СММ - скрытая марковская модель

ПФБД - персональная фонетическая кластерная база данных

РЕ - речевая единица

РК - речевой корпус

ФАР — фонетический анализ речи

ФБД - фонетическая база данных

ЭРЕ - элементарная речевая единица

Введение

Актуальность темы исследований. В современных условиях информационного общества с каждым днем все более актуальным становится использование речевых технологий, таких как, например, распознавание, анализ речи, голосовое управление сложными техническими системами, а также автоматизированная постановка произношения. Данные технологии широко используются при построении справочных и поисковых систем, систем интерактивного обучения иностранным языкам или постановки произношения у глухих и слабослышащих детей, создании разнообразных речевых корпусов, предназначенных как для исследовательских целей (например, сопоставлении различных диалектов), так и для предварительного обучения систем распознавания и синтеза речи.

Одним из основных направлений развития речевых технологий можно считать задачу автоматического распознавания речи (АРР). Ее особенностью является то, что обрабатывается мультимедийная информация. Как известно [1], основной характеристикой подобного вида информации является ее протяженность во времени [2]. Указанное обстоятельство накладывает некоторые ограничения на используемые для решения этой задачи методы и подходы [3-9].

Один из таких подходов основан на использовании искусственных нейронных сетей (ИНС), большой вклад в развитие которых внес Минский М. Л. [10]. При применении ИНС в задаче автоматического распознавания аудио и видео образов предпочтение обычно отдается многослойным сверхточным искусственным нейронным сетям [11]. Они реализуют собой *детерминистский* подход. При использовании таких сетей во входном объекте выделяются так называемые карты признаков. В задаче распознавания ИНС позволяют автоматически настроить систему для эффективного различения набора распознаваемых слов, однако цепочки слов, а также слова, произносимые с разным темпом, нейронные сети идентифицируют плохо.

При быстром произношении слова некоторые фонемы могут пропадать («проглатываться»). Поэтому следующий этап классификации представляет собой выравнивание [12] элементарных единиц тестируемого объекта с каждым из объектов-эталонов. Математическим аппаратом, который позволяет выполнить указанный выше процесс выравнивания, является аппарат скрытых Марковских моделей (СММ) [3]. СММ являются одним из способов получения математической модели (описания свойств) некоторого наблюдаемого случайного процесса. Данный математический аппарат был предложен Баумом Л.Е. и его коллегами [13 - 15] и впервые был использован в задаче распознавания речи в работах Бейкера Дж.К. [16, 17] и Лоуэрра Б.Т. [18]. Дальнейшее свое развитие применительно к задаче АРР аппарат СММ получил в работах многих ученых. Так Липорак Л.Р. [19] предложил обобщение метода оценки параметров непрерывных СММ на случай использования смесей логарифмически вогнутых и/или эллиптически симметричных распределений. Жуань Б. и др. [20] получили формулы переоценки параметров СММ для случая аддитивной суммы гауссовых распределений. Левинсоном С.Е. [21] была предложена параметрическая модель длительности состояний. Существенный вклад в развитие систем распознавания на основе скрытых марковских моделей внес Рабинер Л., впервые опубликовавший описание масштабированного прямого обратного метода обучения СММ [3]. Под его руководством было создано множество автоматизированных систем голосового сервиса, которые смогли заменить живых операторов.

Следующим практическим применением автоматического анализа речи является задача формирования и обновления речевых корпусов (РК) — структурированных совокупностей речевых фрагментов, обеспеченных программными средствами доступа к ним. Такие корпуса находят широкое применение, прежде всего, в системах анализа и синтеза речи [22]. Здесь следует упомянуть системы идентификации и верификации говорящего по голосу [23 - 28], идентификации психофизического и эмоционального состояния диктора, а также обучающие системы [29 — 38]. Также РК составляют основу

автоматизированных систем сбора и хранения речевых сообщений, поиска и выдачи записанных речевых сообщений по запросу.

Большинство современных РК сопровождается транскрипцией речевых единиц (РЕ), т.е. их описанием через последовательность фонем, аллофонов, дифонов и т.д. Основной проблемой при формировании РК является вариативность речи дикторов, ее составляющих. В связи с этим до настоящего времени не создано ни одной сколько-нибудь эффективной системы автоматического выделения списка РЕ из слитной речи. Например, в информационной системе *Nuance speech recognition* - 8,0 все основные операции над РК, связанные с транскрибированием, выполняются «вручную». Это весьма трудоемкая и дорогостоящая процедура, особенно при учете очевидной необходимости периодического обновления или модернизации существующих баз данных.

Следующим направлением применения автоматического анализа речи является задача автоматизированной постановки произношения, которая предполагает использование компьютеров или других специализированных вычислительных устройств для помощи в обучении произношению. Обучение произношению при помощи компьютерных систем проводится как на уровне отдельных звуков, так и на более высоких уровнях (отдельные слова, фразы и даже небольшие диалоги).

Здесь разработано множество подходов [29 — 38], суть большинства которых заключается в сравнении речи обучаемого с некоторым эталоном. Чем больше сходство речи с эталоном, тем лучше считается произношение. В частности, Росситром Д. [29], была разработана система, использующая коэффициенты линейного предсказания для приближенного отображения формы речевого тракта. Одной из лучших на сегодняшний день является система «Видимая речь III», разработанная фирмой IBM [30]. Кьюли Д. и др. [31 - 33] разработали систему ISTRA, которая работает на принципах, схожих с системой «Видимая речь», но имеет плату распознавания изолированных слов с настройкой на диктора. Среди отечественных ученых можно отметить работу Аграновского

А.В. [38], который разработал систему «Контроль произношения». Данная система постановки произношения также проводит сравнение речи диктора с некоторым эталоном.

Однако, как было выяснено Нери А. [39], основной недостаток систем, построенных по принципу сравнения с эталоном, состоит в том, что даже при очень хорошем произношении речь обучаемого и эталон могут иметь совершенно разные спектры или формы во времени. Кроме того, получаемые результаты достаточно трудно интерпретировать, поскольку нет простого соответствия между артикуляционными движениями и отображаемыми результатами.

Перечисленные выше направления автоматического анализа и распознавания речи в настоящее время реализуются при помощи теории распознавания образов [40]. В рамках данной теории решаются следующие основные задачи: классификация, распознавание и группирование образов. Основное различие между данными задачами заключается в доступной априорной информации.

Задача классификации (распознавания без обучения) заключается в определении по описанию объекта того класса, к которому он принадлежит. При этом решающие правила считаются известными. В рамках дискриминантного [41, 42] подхода это означает, что известны поверхности, разделяющие классы в пространстве признаков, так что для любого объекта, представленного точкой в этом пространстве, можно определить, в какой области он расположен. В рамках синтаксического (лингвистического) подхода [43 - 46] эта задача соответствует ситуации, в которой известны грамматики для соответствующих классов и требуется провести грамматический разбор: определить, является ли структурное описание объекта предложением, синтаксически правильным по отношению к какой- либо из этих грамматик.

Задача распознавания (обучения с учителем) заключается в построении решающих правил, которые считались известными в задаче классификации. В качестве исходной информации здесь выступает обучающая выборка. Каждый элемент выборки представляет собой описание объекта и соответствующий этому

объекту класс. В дискриминантном подходе задача распознавания сводится к построению поверхностей в пространстве признаков, разделяющих заданные в обучающей выборке множества точек. В синтаксическом методе обучения эта задача превращается в задачу обучения грамматикам, т.е. восстановлению грамматик по заданным наборам правильно и неправильно построенных предложений. Решение задачи распознавания должно быть таковым, чтобы обеспечить наиболее высокое качество дальнейшей классификации неизвестных объектов.

Задача группирования (кластеризации) заключается в определении пространства классов, которое требуется сформировать, опираясь на заданный набор образов, не разбитый на классы в отличие от задачи распознавания с учителем. Одной из первых работ, посвященной данной проблеме, была работа Тайрона Р.К. [47].

Формирование классов в задаче группирования соответствует разбиению исходного множества образов на подмножества согласно некоторому критерию качества. Критерий качества группирования должен отвечать на вопросы: почему нельзя объединить все объекты в один класс, или, напротив, ввести для каждого объекта собственный класс? Чем хуже такие разбиения некоторого разбиения с промежуточным числом классов?

Для ответа на эти вопросы необходимо определить понятие близости или сходства образов, поскольку требуется, чтобы подмножества, на которые производится разбиение, включали в себя объекты в некотором смысле более похожие на объекты того же подмножества, чем на объекты, отнесенные к другим подмножествам. В дискриминантом подходе близость объектов трактуется как расстояние между соответствующими точками в пространстве, а группирование — как выделение кластеров — компактно расположенных наборов точек. В связи с этим в рамках дискриминантного подхода задача группирования часто называется задачей кластеризации.

В настоящее время наиболее распространенным подходом при решении перечисленных выше задач анализа и распознавания речи является

статистический (байесовский) подход [48]. В его рамках РЕ представляются гауссовой моделью сигналов и моделируются набором классов. Подобный подход имеет ряд существенных недостатков, таких как невысокая точность и надежность. Для устранения указанных недостатков проф. Савченко В.В. была разработана новая информационная теория восприятия речи (ИТВР) [49], основной которой служит критерий минимального информационного рассогласования (МИР) [50] и кластерная модель речевых единиц. В настоящее время ИТВР можно считать одной из наиболее перспективных ветвей развития акустической теории речеобразования [59 - 69].

Информационная теория восприятия речи в своей идеологии следует в русле гипотезы Л.А. Варшавского и И.М. Литвака о том, что качество звуков зависит, в основном, от уровня соотношений мощности в разных спектральных полосах, а форманты (максимумы в спектре) являются лишь доступным способом достижения необходимых межполосных соотношений. Проблема состоит в том, что особенности частотных спектров сильно варьируются (изменяются произвольным образом) не только от одной речевой единицы к другой, но и от одного диктора к другому в пределах одной и той же речевой единицы и даже в пределах одного диктора в зависимости от времени суток, его эмоционального состояния и других второстепенных факторов. Данный эффект в теории обработки речи известен как проблема вариативности устной речи. Информационная теория восприятия речи предлагает строгий подход к ее решению — на основе теоретико-информационного определения самого понятия «фонема».

Можно утверждать, что одноименные реализации $x_{r,j}$, $j=\overline{1,J_r}$ J_r »1 сознании человека группируются в соответствующие классы или образы речевых единиц $X_r = \{x_{r;}\}$, $r=\overline{1,R}$ вокруг некоторого центра - эталонной метки данного образа [70, 71]. В информационной теории восприятия речи указанные эталоны определяются в строгом теоретико - информационном смысле [72]: речевая метка $x_r^* \subset X_r$ образуется как информационный центр-эталон г-го речевого образа, если в пределах множества X_r она характеризуется минимальной суммой

информационных рассогласований (ИР) по Кульбаку-Лейблеру относительно всех других его меток-реализаций, $x_{r,j}$, $j=\overline{1,J_r}$т. е.

$$x_r^* = x_{rv} : J^{-1}\sum_{j=1}^{J} p\left(\frac{x_{rj}}{x_{rv}}\right) = \min_{i \le J} J^{-1} J^{-1}\sum_{j=1}^{J} p\left(\frac{x_{rj}}{x_{rv}}\right) = p_r^* \quad (0.1)$$

Здесь *p*(*) - информационное рассогласование по Кульбаку-Лейблеру [50].

Таким образом, именно в понятии информационного центра (ИЦ) *r*-го множества реализаций X_r дается наиболее информативное описание свойств соответствующей речевой единицы. Само же множество формируется путем разбиения анализируемого речевого сигнала *X(t)* на ряд последовательных сегментов данных x(t) длительностью 10 — 15 мс [73]. Именно на таких временных интервалах речевой сигнал можно считать стационарным. После этого каждый такой парциальный сигнал рассматривается в пределах конечного списка фонем $\{X_r\}$ и отождествляется с той X_v из них, которой отвечает минимум информационного рассогласования между вектором x(t) и соответствующим эталоном x_v^*, $v \le R$. На рис. 0.1 показана иллюстрация

сформулированного выше определения кластерной модели речевых единиц.

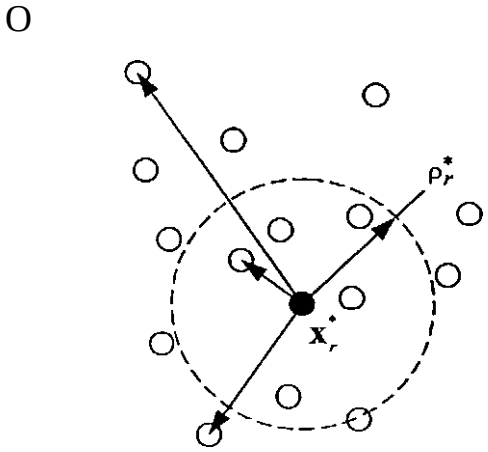


Рис. 0.1 — Кластерная модель

Показано [51, 52], что при несущественных ограничениях данная кластерная модель РЕ охватывает фонетический (звуковой) строй национального языка. Ее применение в задачах распознавания и анализа речи позволило существенно сократить вычислительную сложность используемых для этого алгоритмов (по сравнению с СММ и ИНС) и повысить точность их работы.

Кроме того, кластерная модель позволяет существенно снизить требования к объему обучающей выборки в задаче автоматического распознавания речи. Если скрытой марковской модели для уверенного обучения требуется не менее 20 реализаций каждого слова из распознаваемого словаря, то при использовании кластерной модели для обучения достаточно всего нескольких (в идеальном случае одного) эталонных реализаций каждого слова. Это позволяет не только существенно снизить требования к объему обучающей выборки, но и сократить временные затраты на обучение системы распознавания речи.

Исследованию и практическому применению рассмотренной кластерной модели речевых единиц посвящена данная диссертационная работа.

Объект исследования. Объектом исследования в диссертационной работе является устная речь, ее математические модели и методы обработки.

Предмет исследования. Предметом исследования в диссертационной работе выступает кластерная модель элементарных речевых единиц (ЭРЕ) и критерий минимального информационного рассогласования.

Цель диссертационной работы. Целью диссертационной работы является разработка эффективных (по точности и скорости вычислений) алгоритмов автоматического анализа, распознавания и обучения речи на основе кластерных моделей элементарных речевых единиц.

Задачи исследования. Для достижения поставленной цели в ходе выполнения диссертационной работы были решены следующие основные задачи:

1. Разработка алгоритма распознавания речи на основе кластерной модели элементарных речевых единиц и исследование его эффективности в задаче распознавания изолированных слов.

2. Экспериментальная оценка выигрыша в скорости вычислений в задаче распознавания речи с использованием кластерной модели элементарных речевых единиц.
3. Разработка адаптивной кластерной модели элементарных речевых единиц и ее экспериментальное исследование в задаче фонетического анализа речи.
4. Разработка новой технологии информационного обеспечения с использованием кластерной модели речевых единиц для решения задачи обучения речи и постановки произношения.

Методы исследования. Для решения поставленных задач в работе используются методы теории информации, теории вероятностей, математической статистики.

Научная новизна работы состоит в следующем:

1. Разработан новый алгоритм распознавания речи, основным отличием которого от существующих аналогов является применение кластерной модели элементарных речевых единиц на основе критерия минимального информационного рассогласования, что позволило повысить эффективность распознавания речи.
2. Предложена кластерная модель элементарных речевых единиц, направленная на решение проблемы вариативности речи, основное отличие которой от существующих аналогов состоит в возможности адаптации (автоматического наполнения) в процессе анализа речи.
3. На основе предложенной адаптивной кластерной модели разработана новая технология информационного обеспечения с визуализацией данных в задаче обучения речи и постановки произношения. Ее основным отличием от существующих аналогов является применение двух разновидностей кластерной модели речевых единиц: фонетической и морфологической.

Практическая ценность диссертации обусловлена рядом обстоятельств, а именно:

1. Разработанный алгоритм распознавания речи может применяться в системах автоматического распознавания речи с большим словарем или в условиях

ограниченных вычислительных ресурсов (например, в сотовых телефонах или других встраиваемых платформах).

2. Получены экспериментальные оценки выигрыша по эффективности алгоритма автоматического распознавания речи на основе кластерной модели элементарных речевых единиц по сравнению с распространенными на практике методами на основе скрытых марковских моделей.
3. Предложенная адаптивная кластерная модель элементарных речевых единиц может применяться при создании новых и обновлении существующих фонетических баз данных, которые, в свою очередь, широко используются в поисковых и справочных системах, системах голосового управления и т.д. 4. Разработанная технология информационного обеспечения задачи обучения речи и постановки произношения может быть использована в процессе реабилитации детей-инвалидов с частичной или полной потерей слуха, а также при изучении иностранных языков.

Результаты внедрения. Полученные в диссертации результаты, в частности, адаптивная кластерная модель ЭРЕ, были использованы в итоговом отчете по проекту РФФИ № 07-07-12042-офи «Разработка и создание информационной теории восприятия речи на основе критерия минимума информационного рассогласования», а также в итоговом отчете по проекту РГНФ №09-06-12125в «Разработка информационной системы обучения речи глухонемых и слабослышащих». Созданная информационная система прошла апробацию в городской школе-интернате для глухих и слабослышащих детей и подростков г. Нижнего Новгорода.

Решением Ученого совета результаты диссертационной работы внедрены в учебный процесс НГЛУ им. Н.А. Добролюбова. Кроме того, результаты диссертационной работы внедрены в учебный процесс НГТУ им. Р.Е. Алексеева по дисциплине «Теория сигналов» по темам «Анализ и распознавание речи» и «Адаптивная обработка случайных процессов» в виде компьютерной программы, реализующей разработанные алгоритмы.

Результаты диссертационной работы также внедрены в инновационный проект № 7630р/10278 «Разработка фонетического детектора лжи» в ООО «Системы речевых коммуникаций». Разработка ведется при поддержке Фонда содействия развитию малых форм предприятий в научно-технической сфере.

Основные положения, выносимые на защиту.

1. Алгоритм распознавания речи с нормировкой элементарных речевых единиц и ограничением величины решающей статистики на основе кластерной модели элементарных речевых единиц.
2. Адаптивная кластерная модель элементарных речевых единиц на основе критерия минимального информационного рассогласования, формирование которой происходит в процессе анализа устной речи.
3. Технология информационного обеспечения при обучении речи и постановке произношения, ключевым звеном которой является применение двух разновидностей кластерной модели речевых единиц (фонетической и морфологической), когда речь обучаемого сравнивается не с одним, а с целым набором эталонов.

Структура и объем работы. Диссертационная работа включает введение, четыре главы, заключение, список используемой литературы и приложения. Вся работа изложена на 103 страницы текста, включающих в себя 38 рисуноков, 5 таблиц. Количество библиографических ссылок - 120.

Первая глава **«Задача автоматического распознавания речи»** является обзорной. В ней рассмотрены известные в мировой литературе методы автоматического распознавания речи, в том числе на основе аппарата скрытых марковских моделей. Также в главе кратко рассмотрен критерий минимального информационного рассогласования. Введено строгое определение кластерной модели элементарных речевых единиц. Приведено описание разработанного на ее основе алгоритма автоматического распознавания речи и результатов его математического моделирования. Показано, что применение кластерной модели элементарных речевых единиц в решении задачи распознавания речи позволяет существенно сократить объем необходимых вычислений.

Вторая глава **«Разработка адаптивной кластерной модели речевых единиц»** посвящена созданию алгоритма формирования кластерной модели элементарных речевых единиц из потока слитной речи на основе метода переменного дерева. Особенностью получаемой таким образом модели является возможность ее обновления и адаптации к изменяющемуся фонетическому составу речи. Приведены результаты экспериментальных исследований и оценка эффективности разработанной адаптивной кластерной модели элементарных речевых единиц.

В третьей главе **«Результаты экспериментальных исследований алгоритма фонетического анализа речи на основе адаптивной кластерной модели»** рассмотрен лабораторный образец информационной системы фонетического анализа слитной речи разработанной на основе предложенной адаптивной кластерной модели. Также в главе приводятся **результаты** полевых испытаний лабораторного образца информационной системы фонетического анализа слитной речи на речевом материале, полученном от дикторов нижегородского языкового ареала. Предложен информационный критерий оптимизации фонетической кластерной базы данных, получаемой на основе адаптивной кластерной модели. Приведены **результаты** применения данного критерия в задаче выбора оптимального диктора. Предложен способ повышения точности работы алгоритма автоматического распознавания речи.

Четвертая глава **«Применение адаптивной кластерной модели в новой технологии информационного обеспечения при обучении речи слабослышащих»** посвящена возможностям разработанных алгоритмов анализа и распознавания речи при создании новой технологии информационного обеспечения при обучение речи и постановке произношения. Приведено описание лабораторного образца информационной системы обучения речи глухонемых и слабослышащих, разработанного для апробации и исследования предложенной технологии. Даны результаты полевых испытаний.

В **заключении** представлены основные результаты диссертационной работы и сделан ряд выводов в отношении эффективности использования кластерной модели речевых единиц в задачах анализа и распознавания речи.

ГЛАВА 1. Задача автоматического распознавания речи

1.1. Постановка задачи автоматического распознавания речи. Теоретико-вероятностный подход

В настоящее время речевые технологии доступны для узкого, но интересного круга задач. Эти технологии позволяют машинам правильно реагировать на человеческий голос и предоставлять различные сервисы. Кроме использования в колл-центрах распознавание речи широко применяется в задачах управления технологическими процессами, системах безопасности, системах управления различными бытовыми приборами и программным обеспечением.

В основе технологий голосового управления и человеко-машинного общения лежат различные алгоритмы и методы АРР. В рамках теории распознавания образов задачу АРР в общем виде можно представить следующим образом:

Пусть мы имеем набор эталонной моделей РЕ $\{x_j\}$,r $=\overline{1,R}$ (это могут быть фонемы, аллофоны, целые слова и т.д.) для которых определена некоторая мера сходства $\rho(x_p x_2)$. Тогда для того, чтобы распознать к какой эталонной модели относится анализируемая речевая единица *X*, необходимо решить задачу многоальтернативной проверки гипотез в виде

$$W_v(X): \qquad p_{x,v}=(X,x_r)\big|_{v=r} \to min\, r=\overline{1,R} \tag{1.1}$$

На рис. 1.1 показана блок-схема решения задачи (1.1) для случая распознавания изолированных слов.

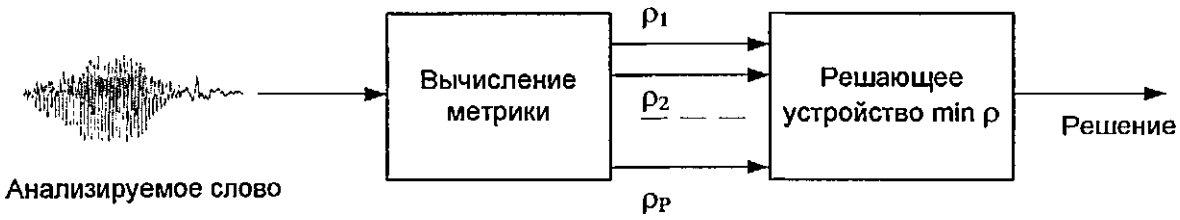


Рис. 1.1 - Блок-схема устройства распознавания речи

Для решения задачи АРР, сформулированной в виде (1.1) на практике необходимо задать меру сходства $\rho(x_1,x_2)$ для сравнения между собой РЕ. Здесь было предложено множество различных вариантов [4, 116, 117]. Однако наиболее распространенным из них на сегодняшний день является применение вероятностной меры сходства. Она применяется в различных стохастических моделях, наиболее широко распространенной из которых применительно к решению задачи АРР является скрытая марковская модель. СММ была впервые предложена Баумом и его коллегами в 60-х гг. 20-го века [13 — 15]. Однако в задаче распознавания речи данный подход начал применяться только вначале 80-х гг.

Скрытая марковская модель представляет дважды стохастический процесс и позволяет моделировать существенные изменения речевого сигнала (и результирующие спектральные признаки) таким же образом, как и структуру разговорного языка в завершенном и стройном аппарате статистического моделирования. Как известно, речевой сигнал может в значительной степени изменяться. Когда человек дважды произносит одно и то же слово, акустические сигналы не являются одинаковыми (они могут очень сильно различаться), несмотря на лежащую в их основе лингвистическую структуру. В зависимости от произношения, синтаксиса и грамматики, данное слово может (или не может) оставаться тем же самым. Аппарат СММ - это вероятностная мера, которая использует марковскую цепь для представления лингвистической структуры и множества распределений вероятностей для учета вариативности реализаций фонем в предложении. При этом задача выбора гипотезы при отнесении анализируемого сигнала к одному из классов может быть сформулирована следующим образом

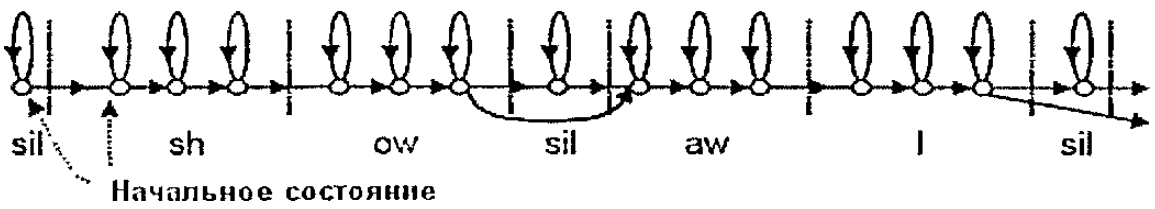
$$W_v(X): \qquad L_v(X) = max\, L_r(X), r = \overline{1,R} \tag{1.2}$$

где $L_r(X)$ - функция правдоподобия для r-ой скрытой марковской модели.

Имея набор известных предложений, представленных набором реализаций интересующих нас слов (называемых обучающим множеством), можно использовать эффективный алгоритм, называемый алгоритмом Баума - Уэлча

[55], для получения «наилучшего» множества параметров, которые определяют соответствующую модель или модели. Оценка параметров, определяющих модель, эквивалентна тренировке или обучению. Полученная модель затем используется для определения вероятности того, что неизвестный входной сигнал является одной из реализаций данного слова (или слов) представляемых моделью. Вероятностная мера, выдаваемая скрытой Марковской моделью, является неотъемлемой частью систем распознавания речи, использующих статистический подход к распознаванию образов и опирающихся на байесовскую теорию принятия решений. Методология СММ является большим шагом вперед от простого распознавания образов и акустико-фонетических методов, применявшихся в ранних системах автоматического распознавания речи.

СММ, начавшись с вероятностной меры, хорошо поддавались объединению в обширный аппарат расшифровывания речи, который включил в себя и модель языка. Использование автоматных грамматик в распознавании слитной речи с большим словарем позволило усовершенствовать цепи Маркова, когда СММ используется для учета структуры языка, а не только на уровне, на котором учитывается взаимодействие между артикуляцией и произношением. Хотя эти модели (для различных уровней языковых ограничений) являлись лишь хорошими начальными приближениями реального речевого сигнала, они были вычислительно эффективны и часто приводили к хорошим результатам. Объединение Марковских моделей (с их преимуществом в статистической устойчивости, особенно в обработке акустической вариативности) и сетей с конечным числом состояний (с их поисковой и вычислительной эффективностью, особенно при выборе гипотезы, соответствующей последовательности слов) было важным, хотя и неожиданным, развитием методов распознавания в середине 80-х годов.



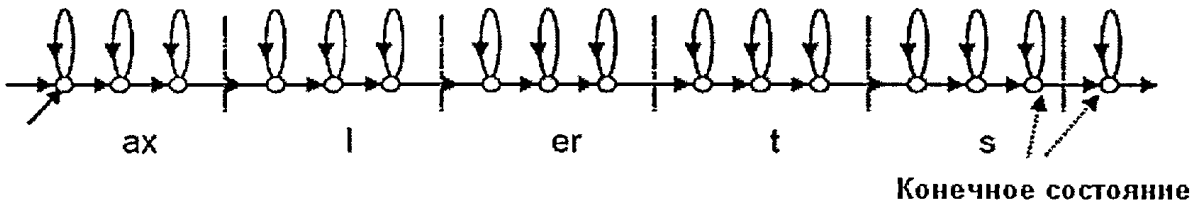


Рис. 1.2 - Скрытая марковская модель фразы «show all alerts»

На рис. 1.2 показана составная модель с конечными состояниями для предложения «show all alerts», построенная из различных контекстно-зависимых моделей частей слова, представляющие соответствующие фонем подобные речевые единицы (включая единицы для тишины, которая может присутствовать в начале и в конце предложения, а также в конце любого слова в предложении, за которым может следовать пауза в произношении). Граф с конечными состояниями реализуется как цепь Маркова для расчета вероятности на основе последовательности наблюдений (спектральное представление во времени) неизвестного входного сигнала. Следует отметить, что каждая вершина в графе связана с распределением вероятности, которое учитывает вариативность реализаций соответствующих фонем подобных звуков. Вероятность того, что входная последовательность была сгенерирована сетью с конечными состояниями, представленной данной моделью, рассчитывается как сумма локальных вероятностей (относительно элементарных единиц составной модели) после применения процедуры динамического программирования, которая позволяет максимизировать величину совпадения между обозначенными элементами и соответствующими частями входной последовательности наблюдений (даже для моделей с неправильными последовательностями слов). В любой момент времени мы имеем множество предполагаемых речевых единиц и устанавливаем соответствие какой-либо из них определенному звуку на основе максимума значения вероятности (или оценке совпадения). Число рассматриваемых единиц и, следовательно, число путей для поиска может достигать астрономических размеров.

Наиболее существенным недостатком скрытых марковских моделей является тот факт, что они не умеют выстраивать разницу между

распознаваемыми классами, не умеют находить отличительные характеристики. То есть СММ не обладает различающей способностью: алгоритм обучения только максимизирует отклик каждой модели на свои классы, но не минимизирует отклик на другие классы; не выделяются ключевые признаки, отличающие один класс от другого. Таким образом, похожие классы могут оказаться слабо различимыми и при увеличении объёма базы или использования в более широких условиях СММ могут оказаться ненадёжными [56].

Еще одна проблема - это проблема точности СММ (об этом см., напр., работу [9]). Особенно актуальной она становится в задачах распознавания из больших словарей [57]. Здесь по каждому отдельному классу для обучения скрытой марковской модели требуются десятки, и даже сотни независимых образцов. Проблема точности СММ резко обостряется на практике еще и малыми выборками наблюдений, а также связанной с ней проблемой неоднородности речи во времени. Так широко известная система распознавания речи *Nuance Speech Recognition System* [58] обеспечивает распознавание более чем 10 тысяч слов. Однако для обеспечения требуемой точности распознавания словарь системы имеет иерархическую структуру. При этом на каждом уровне иерархии используется словарь небольшого объема (всего около 50 слов).

При распознавании больших словарей при помощи аппарата СММ также возникает проблема вычислительной сложности применяемых алгоритмов. На практике уже при объеме словаря, превышающем 1000 слов, возникают проблемы с распознаванием речи в реальном времени. Поэтому разработка алгоритмов и методов автоматического распознавания речи, позволяющих существенно снизить требования к необходимому объему вычислений, является на сегодняшний день весьма актуальной задачей.

В последнее время стало популярным использование так называемых гибридных моделей, объединяющих в себе преимущества СММ и ИНС. Благодаря синтезу двух подходов удалось ослабить недостатки, присущие каждому из них, однако они не могут быть устранены полностью.

Для преодоления указанных выше недостатков, присущих современным методам автоматического распознавания речи, несомненный интерес представляет использование критерия минимального информационного рассогласования [71, 76]. Подробное рассмотрение данного критерия приведено в следующем разделе.

1.2. Теоретико-информационный подход. Критерий минимального информационного рассогласования

Задача статистической классификации в своей стандартной формулировке обычно сводится к проверке Л гипотез о законе распределения выборки-сигнала [71]:

$$W_r: \quad P_x = P_r, r = \overline{1, R} \tag{1.3}$$

Здесь P_r - r-я альтернатива распределения, предполагаемая точно заданной; при этом множество альтернатив $\{P_r\}$ исчерпывает собой все допустимое многообразие законов распределения выборки *X*. Оптимальное в байесовском смысле решение в пользу одной из гипотез W_v, $v \le R$ основывается в общем случае на классическом критерии максимального правдоподобия:

$$W_v(X): \quad L_v(X) = max \;\; L_r(X) \tag{1.4}$$

где $L_r(X)$ — функция правдоподобия для r-ой гипотезы W_r. Предположим, что наша выборка (сигнал) *X* составлена из *M* независимых векторов (столбцов) $x_m, m = \overline{1, M}$ размерностью $n > 1$ каждый и подчиняется центрированному гауссову (нормальному) распределению. В таком случае можно записать

$$\ln L_r(X) = -0{,}5\, M^{-1}\left[\ln\left|K_r\right| + tr\left(S_x K_r^{-1}\right) + n \times \ln(2\pi)\right], \qquad r = \overline{1; R} \tag{1.5}$$

где K_r - (n×n)-матрица автоковариаций (АКМ) для r-ой альтернативы закона распределения P_r , а $S_x = M^{-1} \sum_{m=1}^{M} x_m x_m^T$ - выборочная оценка АКМ (здесь символами tr(o) и |o| обозначены соответственно след и определитель квадратной матрицы).

Выражения (1.4) и (1.5) в совокупности и определяют для заданных условий вид оптимального алгоритма обработки сигналов в задаче распознавания образов [90]. Видно, что здесь в качестве базы априорных данных используется набор из R альтернативных вариантов АКМ $\{K_r\}$. А сама обработка сводится к R параллельным вычислениям функции правдоподобия (1.5) на заданном множестве АКМ с последующим сравнением полученных результатов по величине и выбором максимального из них. Это стандартная поисковая процедура прямого перебора альтернатив. Причем в задачах с априорной неопределенностью в (1.5) используются статистические оценки альтернативных АКМ по Л классифицированным выборкам предварительных наблюдений.

Проблемы возникают, однако, в случае отсутствия заранее классифицированных выборок, т.е. в задачах статистической классификации «без учителя». Например, при анализе фонетического состава речи и статистические характеристики ЭРЕ, и их суммарное число Я зависят от особенностей голосового аппарата каждого конкретного диктора. Выход в указанных ситуациях может быть найден в реализации поисковой процедуры (1.4), (1.5) по методу дерева [74], но с переменным числом вершин 2,3, ..., или по методу «переменного» дерева. Указанный подход требует упорядочения набора альтернативных образов $\{P_r\}$ (или $\{K_r\}$) в некотором смысле или в некоторой метрике. Здесь может быть использована, например, классическая евклидова метрика, вычисляемая в отношении каждого варианта АКМ. В таком случае обработка выборки X будет состоять в поиске альтернативы K_y с минимальной нормой $\| S_x - K_v \|$ на множестве $v \le R$.

Очевидно, однако, что предлагаемый алгоритм будет заведомо проигрывать по своей эффективности оптимальному (1.4), (1.5). Поэтому предпочтительнее представляется иной подход [71], основанный на упорядочении множества

альтернативных распределений $\{P_r\}$ по величине их взаимного информационного рассогласования (ВИР) в метрике Кульбака-Лейблера [50]. Несмотря на то, что информационное рассогласование не является метрикой в строгом смысле, предлагаемый подход имеет строгое теоретическое обоснование [75].

Утверждение 1. В условиях введенных выше ограничений в задаче распознавания образов (1.3) критерий максимального правдоподобия в формулировке (1-4), (1-5) реализует критерий минимального информационного рассогласования на множестве упорядоченных пар *n*-мерных гауссовых распределений (P_x,P_r), $r = \overline{1;R}$, с АКМ, равными S_x и K_r соответственно.

Доказательство прямо следует из сопоставления набора решающих статистик (1.4) с известным выражением для ВИР в гауссовом случае [50]

$$J\left(\frac{P_x}{P_r}\right)=0{,}5\left[-\ln\left|K_r\right|-\ln\left|S_x\right|+tr\left(S_x K_r^{-1}\right)-n\right]\geq 0, \qquad (1.6)$$

которая равна нулю лишь при условии эквивалентности распределений, когда выполняется равенство $S_x = K_r$.

В работе [75] также было показано, что в асимптотике, когда $n\to\infty$, при гауссовом распределении сигнала $P(X_r)$ с АКМ K_r ленточной структуры выражение для оптимальной решающей статистики сводится к виду

$$p_{x,r}=\frac{1}{F}\sum_{f=1}^{F}\left(\frac{G_x(f)}{G_r(f)}+\ln\frac{G_r(f)}{G_x(f)}\right)-1\to min\big|_{r=\overline{1,R}} \qquad (1.7)$$

Здесь $G_x(f)$ — выборочная оценка спектральной плотности мощности (СПМ) сигнала X в функции дискретной частоты f; $G_r(f)$ - СПМ r-го сигнала из словаря эталонов; F — верхняя граница частотного диапазона сигнала или используемого канала связи. Это известная формулировка критерия МИР на основе АР-модели речевого сигнала:

$$x(n)=\sum_{i=1}^{p} a(i)\,x(n-i)+\varepsilon(n) \qquad (1.8)$$

Здесь x(n) — значение n -го отсчета речевого сигнала, **a** = {a(i)} — вектор его АР-коэффициентов, P — порядок АР-модели, а $\varepsilon(n)$ - порождающий процесс типа белого гауссова шума (БГШ) с нулевым значением математического ожидания и фиксированной дисперсией σ^2.

При дополнительном условии нормировки АР-модели сигналов типа ЭРЕ по дисперсиям их порождающего шума [76] второе слагаемое в правой части (1.4) оказывается тождественно равным нулю, и выражение для решающей статистики МИР (1.7) приобретает предельно простой вид

$$p_{x,r} = \frac{1}{F}\sum_{f=1}^{F} \frac{\left|1 + \sum_{m=1}^{P} a_r(m)\exp\left(-\frac{j\pi mf}{F}\right)\right|^2}{\left|1 + \sum_{m=1}^{P} a_x(m)\exp\left(-\frac{j\pi mf}{F}\right)\right|^2} - 1, \qquad (1.9)$$

Это стандартная формулировка метода обеляющего фильтра (МОФ) в задаче АРР на основе выборочной оценки ВИР (1.6) между сигналом *X* на входе и *r*-м сигналом из словаря в частной области. Преимуществом данной интерпретации критерия МИР является, прежде всего, возможность его эффективной реализации в адаптивном варианте на основе быстрых вычислительных процедур авторегрессионного анализа, таких как метод Берга и др. [78]. Аналогичным образом можно записать выражение для решающей статистики во временной области [76]:

$$W_v(X): \quad \rho_{r,x} = \sigma_r^2(X) - 1\big|_{r=v} = min, \qquad (1.10)$$

где $\sigma_r^2(X)$ - дисперсия отклика *r*-го обеляющего фильтра на входной сигнал *X*.

Таким образом, рассмотренная в этом разделе ИТВР [49], основываясь на критерии МИР Кульбака-Лейблера [50] и АР-модели из теории акустической трубы Гельмгольца [59], предоставляет исследователю следующие основные результаты: принцип МИР в частотной области (1.7), метод обеляющего фильтра с нормировкой входных сигналов по дисперсии порождающего шума (1.9), (1.10) и, главное, строгое определение кластерной модели ЭРЕ через множество ее однотипных реализаций, объединенных по критерию МИР вокруг своего «геометрического» центра - ИЦ-эталона (0.1).

Основные преимущества использования кластерной модели ЭРЕ - это сокращение объема обучающей выборки и повышение скорости вычислений. Покажем это на примере сравнения с методами на основе СММ.

На рис. 1.3 показана решетчатая структура вычисления вероятности появления анализируемой последовательности наблюдений для заданной СММ.

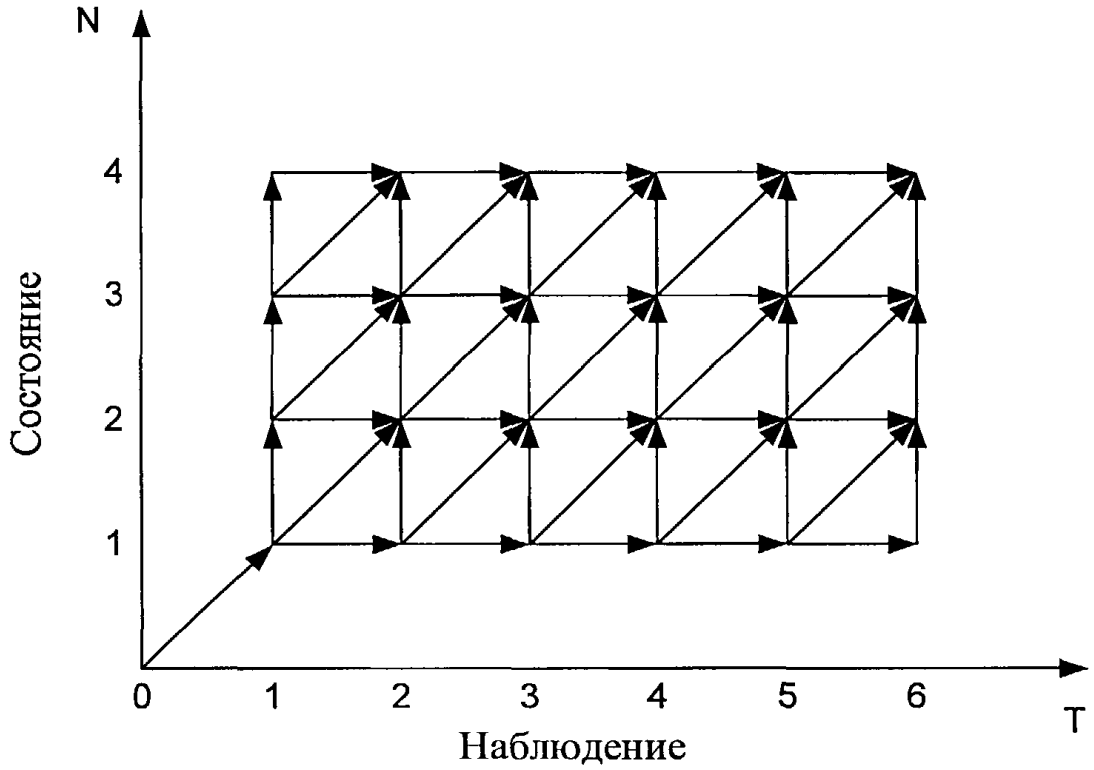


Рис. 1.3- Решетчатая структура вычисления вероятности в СММ

Из рисунка видно, что в общем случае нам необходимо получить значения вероятностей для всех элементов решетки. Полагая, что Я - число состояний СММ, Ь — число сегментов, на которое разбивается входное слово, *n* - длина одного сегмента в отсчетах, *p* - порядок авторегрессионной (АР) модели вектора речевых признаков можно показать, что для словаря объемом *M* слов количество необходимых операций определяется выражением [77]

$$V_1 = \widetilde{R}\, ML(n-p)(p+1) \qquad (1.11)$$

С другой стороны, поскольку при использовании кластерной модели ЭРЕ нам не требуется каждый раз проводить вычисления рассогласований для всех элементов решетки (их можно вычислить заранее), то вычислительная сложность алгоритма распознавания определяется формулой [77]:

$$V_2 = RL(n - p)(p + 1) \qquad (1.12)$$

где R — число четко проговариваемых фонем в языке. Например, при L=100 с, n=80 отсчетов, R=30 и p=20 имеем V_1 - 3780 тыс. элементарных операций за секунду. Для сравнения СММ-методам при M=1000 (соответствует достаточно небольшому словарю), $\widetilde{R}=5$ и той же длине сигнала L=100 сегментов число выполняемых операций составит почти 630 млн. единиц, т. е. увеличится более чем на два порядка.

Представленные результаты в дальнейшем будут использованы при решении задач анализа и распознавания речи, рассмотренных в данной диссертационной работе:

- задача АРР на примере распознавания изолированных слов (раздел 1.3);
- задача формирования фонетической кластерной базы данных (ФБД) (главы 2 и 3);
- задача постановки произношения (глава 4).

1.2. Алгоритм распознавания изолированных слов

Для иллюстрации возможностей и преимуществ кластерной модели ЭРЕ в задаче АРР рассмотрим алгоритм распознавания изолированных слов с учетом их фонемной структуры.

Традиционным элементом обработки сигналов в задаче АРР является процедура их предварительного сегментирования на короткие отрезки данных [79, 80]. В результате задача сводится к поэтапному распознаванию каждого отдельного сигнала. На первом этапе распознаются элементарные речевые единицы типа отдельных фонем. На втором - слова, фразы и даже целые тексты как соответствующим образом структурированные последовательности разных фонем. Таким образом, задача распознавания элементарных речевых единиц является ключевой при автоматическом распознавании речи.

Разбиение слов на фонемы как первый этап АРР является отдельной сложной задачей прикладной лингвистики [82]. Эффективным способом учета тонкой структуры речевого сигнала является метод посегментного анализа, который заключается в последовательной обработке коротких (стационарных) участков слова согласно решающему правилу (1.10). Поскольку любой диктор не в состоянии абсолютно точно, в одном и том же темпе, произнести любое слово даже дважды, сигнал X и эталон X_r должны быть предварительно выровнены по темпу речи. Для этого в задачах АРР применяется известный метод динамического программирования [81]. Воспользуемся им и на основе выражения (1.10) для каждого *1-го* сегмента слова X на входе получим набор минимальных решающих статистик вида

$$\rho_{l,r}(X) = min\left[\rho_{l,r}(X,k), \rho_{l,r}(X,k+1)\right],\ l=\overline{1,L},\quad k=\overline{1,\ N_r} \qquad (1.12)$$

Здесь L - число сегментов, на которые разбивается анализируемое слово X, N_r - аналогичное число стационарных участков в r-м слове-эталоне, а $p_{l,r}$(X,k) — решающая статистика (1.10), вычисленная для *1-го* сегмента слова X и *k-го* участка эталона X_r. Записанный алгоритм можно представить в виде набора из R конечных автоматов по числу анализируемых слов (рис. 1.4). Переход в каждое очередное состояние происходит в них при условии $p_{l,r}$ (X,k_r +1) < $p_{l,r}$ (X, k_r). В противном случае *г*-й конечный автомат остается в том же состоянии, в котором он находился раннее.

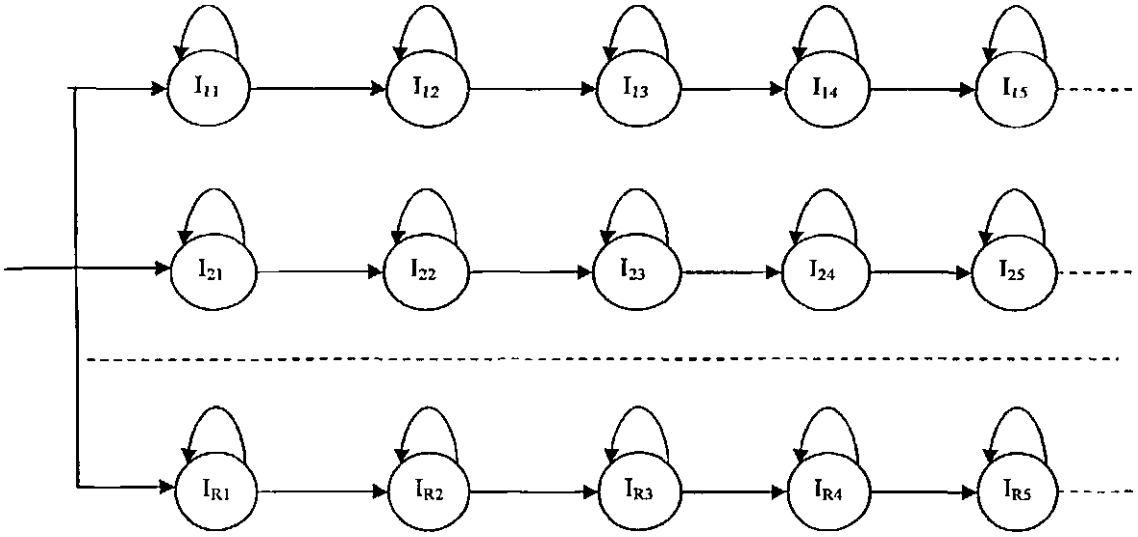


Рис. 1.4 - Представление метода ДП в виде набора конечных автоматов

В предположении о статистической независимости отдельных сегментов между собой величина информационного рассогласования по слову целиком определяется суммарным выражением вида

$$W_v(X):\quad \Lambda_v(X)=\min_r \sum_{l=1}^{L} \rho_{l,r}(X)\big|_{r=v},\quad r=\overline{1,R} \tag{1.13}$$

Распознавание слов методом обеляющего фильтра с сегментированием (1.10)...(1.13) реализуется на основе многоканальной обработки, в которой число каналов R определяется количеством слов-эталонов. При этом в каждом канале используется набор из N_r обеляющих фильтров, настроенных на последовательные стационарные участки (элементарные речевые единицы) соответствующего эталонного слова. Решение (1.13) принимается по критерию минимума суммы решающих статистик (1.12) по всем L сегментам анализируемого слова.

Реализация рассмотренного алгоритма распознавания речи разделяется на два независимых во времени этапа: создание словаря (обучение системы) и собственно распознавание (принятие решения). Во время первого этапа осуществляется формирование базы априорных данных в виде набора векторов авторегрессии $a_{r,k}$ порядка P для каждого r-го слова, разделённого
на N стационарных участков, $r=\overline{1,R}$. Анализируемое слово X разбивается на L сегментов одинаковой длины по N отсчётов каждый. При этом l - й сегмент нормируется по своей дисперсии порождающего шума $\sigma_l^2(X,l)$ в его АР-модели. Дисперсия некомпенсированного остатка l-го сегмента сигнала X на выходе обеляющего фильтра (ОФ), настроенного по k-му участку r-го сигнала

$$\sigma_{l,r}^2(X,k)=\frac{1}{N-P}\sum_{t=1}^{N-P} z_{l,r,k}^2(t) \tag{1.14}$$

определяется $(N-P)$ отсчётами сигнала $z_{l,r,k}^2(t)$ на выходе соответствующего ОФ:

$$z_{i,r,k}(t)=x_l(t)-\sum_{j=1}^{P} a_{r,k}(j)\times x_l(t-j)\big|_{t=\overline{1,N-P}} \tag{1.15}$$

Его коэффициентами служат параметры АР-модели $a_{r,k} = \{a_{r,k}(j)\}$, $j=\overline{1,P}$, соответствующие k-му участку r-го сигнала и взятые с противоположным

знаком. Коэффициенты каждого фильтра определяются рекуррентной процедурой Берга-Левинсона:

$$a_m(i) + c_m a_{m-1}(m-i), \quad i = \overline{1,m};$$

$$c_m = S_{m-1}^{-2} \sum_{l=m}^{N-1} \eta_{m-1}(t) V_{m-1}(t-1),$$

$$S_{m-1}^2 = 0{,}5(n-m)^{-1} \sum_{l=m}^{N-1} \left[\eta_{m-1}^2(t) + v_{m-1}^2(t-1) \right] \qquad (1.16)$$

$$\eta_m(t) = \eta_{m-1}(t) - c_m v_{m-1}(t-1),$$

$$v_m(t) = v_{m-1}(t-1) - c_m \eta_{m-1}(t), \quad t = 0,1,\ldots,N-1,$$

$$\sigma_m^2 = \left(1 - c_m^2\right)\sigma_{m-1}^2, \quad \sigma_0^2 = S_0^2, \quad m = \overline{1,p}.$$

Процедуры вычисления всех векторов $a_{r,k}$ одинаковы и отличаются лишь используемыми в начале процедуры сигналами $x_{r,k}$ (t). Следовательно, коэффициенты каждого ОФ $a_{r,k}$ (j) = a_p (j), $j=\overline{1,P}$. определяются по окончанию рекуррентной процедуры при значении порядка $m = P$.

Таким образом, для вычисления результирующей статистики (1.12) методом ДП производится обработка каждого l-го сегмента сигнала X обеляющими фильтрами порядка P с векторами АР-коэффициентов $a_{r,k}$ и $a_{r,k+1}$ соответственно. Дисперсия некомпенсированного остатка для каждого l-го сегмента входного сигнала на выходе соответствующего ОФ подставляется в (1.12) для получения решающей статистики по каждому сегменту. Решение принимается по признаку минимума значения статистки (1.13).

Экспериментальные исследования алгоритма АРР проводились на речевой базе, состоящей из десяти числительных русского языка: «ноль», ..., «девять». Задачи распознавания речи такого типа относятся к задачам голосового управления автоматическими системами и являются актуальными для практического применения. Частота дискретизации 8 КГц и квантование 8 бит были выбраны из соображений минимизации потока речевых данных, а также для соответствия параметрам передачи и хранения речи, принятым в большинстве стандартов. Для максимального приближения эксперимента к практической ситуации были использованы стандартный канал встроенного звукового процессора бытового персонального компьютера и обычный микрофон.

При формировании словаря эталонов сначала был записан набор фонем в режиме их продолжительного звучания. После этого по каждой фонеме был вычислен вектор АР-коэффициентов с использованием процедуры Берга-Левинсона (1.16) в результате чего была сформирована кластерная модель $Y=q(X_r)$. Затем каждое слово эталонного словаря $\{X_r\}$,$r = \overline{1,R}$ кодировалось с использованием полученной кластерной модели Y (см. рис. 1.5). В результате этого каждый эталон представлял собой набор векторов авторегрессии $a_{r,k}$,$r= \overline{1,R}$.



Рис. 1.5 - Схема процедуры построения кластерной модели эталонного слова

Для создания экспериментальной речевой базы были записаны тем же диктором с одной интонацией по сто реализаций слов каждого из десяти числительных.

Результаты распознавания слов по алгоритму (1.10)... (1.16) с использованием описанной экспериментальной речевой базы показаны в таблице 1.1.

Таблица 1.1 - Вероятность правильного распознавания слов из экспериментальной базы с использованием сформированной кластерной модели

Слово	0	1	2	3	4	5	6	7	8	9
Вероятность распознавания	0,93	0,91	0,85	0,9	0,89	0,94	0,81	0,79	0,95	0,97

В таблице 1.2 приведены результаты распознавания при использовании кластерных моделей, сформированных из других реализаций произнесенных фонем.

Таблица 1.2 — Вероятность правильного распознавания слов из экспериментальной базы для различных кластерных моделей

Слово	**Вероятность распознавания**			
	Кластерная модель 1	**Кластерная модель 2**	**Кластерная модель 3**	**Кластерная модель 4**
0	**0,93**	**0,81**	**0,85**	**0,88**
1	**0,91**	**0,98**	**0,94**	**0,91**
2	**0,85**	**0,95**	**0,86**	**0,97**
3	**0,9**	**0,93**	**0,98**	**0,74**
4	**0,89**	**0,99**	**0,73**	**0,82**
5	**0,94**	**0,9**	**0,81**	**0,79**
6	**0,81**	**0,99**	**0,92**	**0,91**
7	**0,79**	**0,63**	**0,77**	**0,95**
8	**0,95**	**1**	**0,81**	**0,82**
9	**0,97**	**0,98**	**0,98**	**0,96**
Итого	**0,894**	**0,916**	**0,865**	**0,875**

Из таблицы видно, что качество распознавания слов из экспериментальной речевой базы существенно зависит от выбора эталонных реализаций фонем. В связи с этим возникает проблема выбора оптимальной фонетической базы.

Покажем теперь преимущества разработанного алгоритма с использованием кластерной модели ЭРЕ перед методами распознавания, основанными на аппарате СММ. Для этого было проведено исследование зависимости времени распознавания одного слова от объема словаря для алгоритма (1.10)...(1.17) и алгоритма АРР на основе СММ.

Для сравнительного анализа использовалась неэргодическая СММ с непрерывной плотностью вероятности наблюдений [81, 19] с 7 состояниями. Каждое слово экспериментального словаря было представлено отдельной СММ. Обработка данных велась посегментно. Длительность каждого сегмента, на которые разбивалась входная реализация, составляла 20 мс с перекрытием в 10 мс. В качестве векторов-признаков для представления речи использовались кепстральные коэффициенты 12-го порядка, рассчитанные по каждому анализируемому сегменту. Обучение СММ велось при помощи алгоритма Баума-Уэлча [55], а распознавание - при помощи алгоритма Витерби [88, 89].

Речевая база для экспериментальных исследований состояла из 30 числительных русского языка по 100 реализаций каждого слова, записанных в отдельных звуковых файлах. Частота дискретизации записей составляла 8 кГц, 16 бит. Экспериментальные исследования всех алгоритмов проводились на ПК Athlon XP 1400 MHz, 512 Mb RAM. Оба алгоритма работали под управлением Matlab R2006a. Полученные результаты представлены на рис. 1.6.

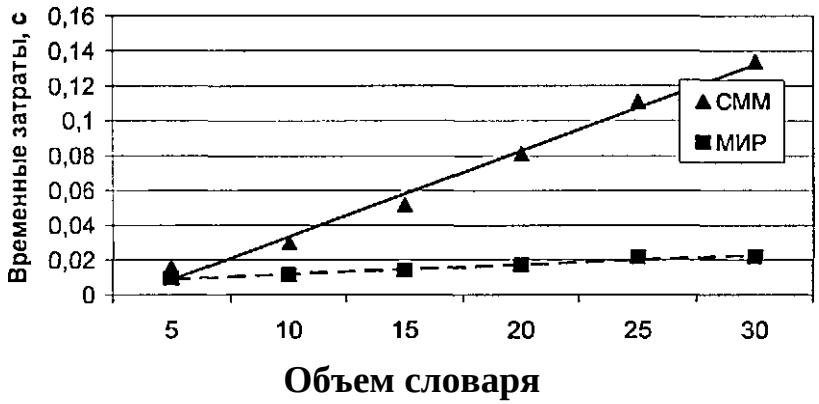


Рис. 1.6 - Сравнение скорости распознавания изолированных слов. Здесь СММ — алгоритм на основе скрытой марковской модели, МИР - алгоритм на основе критерия МИР и кластерной модели ЭРЕ.

Здесь СММ - алгоритм на основе аппарата СММ, МИР - разработанный алгоритм (1.10)...(1.16) на основе кластерной модели ЭРЕ и критерия МИР. Из рисунка видно, что алгоритм распознавания слов на основе кластерной модели ЭРЕ обладает существенно лучшими

характеристиками быстродействия, чем алгоритм на основе скрытых марковских моделей. Сводные результаты сравнения выигрыша в скорости вычислений по отношению к СММ-методам для словаря в 30 слов приведены в таблице 1.3.

Таблица 1.3- Сравнение выигрыша в скорости вычислений

Алгоритм	СММ	МИР
Относительный выигрыш	1	6,05

Из полученных результатов можно сделать вывод, что использование кластерной модели ЭРЕ и свойств решающей статистики МИР в решении задачи АРР позволяет существенно (более чем в 6 раз при объеме словаря в 30 слов) сократить объем необходимых вычислений по сравнению с широко используемыми сегодня методами на основе аппарата скрытых марковских моделей.

Выводы по главе

В главе приводится обзор современных методов и подходов к задаче АРР. Рассмотрены их достоинства и недостатки. Сделан вывод, что наиболее широкое применение на сегодняшний день имеют методы на основе СММ, а также показаны их основные достоинства и недостатки. Для преодоления недостатков существующих методов АРР предложен новый теоретико- информационный подход к задаче распознавания речи. В его основе лежит ИТВР и кластерная модель речевых единиц. ИТВР предоставляет исследователю следующие основные результаты: принцип МИР в частотной области (1.7), метод обеляющего фильтра с нормировкой входных сигналов по дисперсии порождающего шума (1.9), (1.10) и строгое определение элементарных речевых единиц через множество однотипных речевых единиц, объединенных в кластер по критерию МИР вокруг своего «геометрического» центра - ИЦ-эталона (0.1).

Приводятся результаты математического моделирования алгоритма распознавания речи по отдельным словам с учетом их фонемной структуры. Полученные результаты экспериментальных исследований свидетельствуют о том, что эффективность работы алгоритма существенно зависит от выбора эталонных реализаций фонем.

Также показано, что применение кластерной модели элементарных речевых единиц позволяет существенно сократить вычислительные затраты при решении задачи автоматического распознавания речи по сравнению с существующими методами. Данный факт дает возможность широкого использования кластерной модели элементарных речевых единиц в системах автоматического распознавания речи с большим словарем.

ГЛАВА 2. Разработка адаптивной кластерной модели элементарных речевых единиц

2.1. Задача фонетического анализа речи

Известно [93], что задача формирования ЭРЕ (которые могут представлять собой, к примеру, фонемы) является неотъемлемой частью фонетического анализа речи (ФАР). Соответственно работа всей системы автоматической обработки во многом зависит от эффективности ФАР. При этом состав и качество базового словаря, или фонетической базы данных, некоторого фиксированного объема $R < \infty$ во многом определяют достигаемое качество фонетического анализа речи. Поэтому формирование или настройка фонетической базы данных является ключевой для большинства современных методов обработки речевого сигнала. В рамках ИТВР для решения указанной задачи предложена модель переменного дерева и информационный (R+1)-элемент [94].

Информационный (R+1)-элемент - это условный термин, обозначающий устройство или алгоритм для автоматической классификации или распознавания сигнала X в пределах некоторого множества классов-альтернатив $P_r, r = \overline{1,R}$ [74]. В основе его функционирования используется статистический подход и критерий минимального информационного рассогласования распределений в метрике Кульбака-Лейблера. В отличие от других аналогичных алгоритмов с R выходами (R+1)-элемент имеет дополнительный, (R+1)-й выход, который сигнализирует об отказе при распознавании сигналов одновременно от всех заданных альтернатив. Указанная особенность может служить, в свою очередь, основой для построения эффективного алгоритма статистической классификации сигналов «без учителя», когда ни точные распределения-классы $\{P_r\}$, ни их суммарное число R исследователю заранее не известны. Автоматический анализ фонетического

состава речи относится именно к такому кругу задач. Рассмотрим предлагаемый алгоритм более подробно.

Обозначим анализируемый речевой сигнал X вектором его отсчетов x={x,...,x_n}, где *n* — объем выборки. Выделим в нем первые *m* отсчетов из соображений сохранения в них свойства приблизительной стационарности (однородности распределения) сигнала. Например, при стандартной частоте дискретизации в 8 кГц полагают *m* = 100...200 [73]. Используем полученный сегмент данных x_1 = {x,...,x_m} в качестве обучающей выборки X_1 для оценивания АКМ первой ЭРЕ из нашего сигнала: $\widehat{K}_1 = M^{-1}\sum_{m=1}^{M} x_1 x_q^T$.

Соответствующий закон распределения $P_1 = N(\widehat{K}_1)$ - это первая из вершин нашего будущего «дерева». После этого приравняем Я-1 и возьмем второй сегмент выборки для анализа x_2 = {x_{m+1},...,x_{2m}}. Следуя выражению для решающей статистики (1.9), определим для него удельную, т.е. на один отсчет данных [75], величину информационного рассогласования относительно первой ЭРЕ:

$$p(X_2, X_1) = p_{x,r}\big|_{X = X_2} \qquad (2.1)$$

Полученный результат сопоставляется с некоторым пороговым уровнем p_0 допустимой величины рассогласований между разными реализациями одноименных ЭРЕ устной речи:

$$p(X_1, X_2) \geq p_0 \qquad (2.2)$$

При условии выполнения неравенства (2.2) в нашем дереве появится вторая вершина, и вслед за этим приравниваем число его вершин R =2. В противном случае принимается решение об объединении выборок X_1 и X_2 в одну расширенную выборку первой ЭРЕ, а равенство R=1 сохраняется. Это типичная формулировка информационного (R + 1)-элемента [74].

Вычисления по схеме (1.9), (2.1)...(2.2) повторяются циклически для всех последующих сегментов, данных из первоначальной выборки наблюдений X,

причем повторятся «нарастающим итогом» для переменного R=2,3,... В результате получим дерево с некоторым зафиксированным числом вершин R*. Каждая вершина - это код одной из выделенных при анализе ЭРЕ. Чем больше число вершин в построенном дереве для некоторого конкретного диктора, тем богаче с фундаментальной, фонетической точки зрения его речь. В этом выводе и состоит, по-видимому, главный смысл и назначение фонетического анализа речи. Однако здесь же присутствует и очевидная проблема: чрезмерно большое число ЭРЕ в речи диктора - это признак ее нечеткости, или не информативности. Поэтому после выполнения всех рассмотренных выше вычислений рассортируем полученные вершины по объему $\{V_r\}$ их классифицированных выборок на два множества: множество четких ЭРЕ, для которых выполняется условие

$$V_r \geq V_0, \qquad (2.3)$$

и множество нечетких, сомнительных ЭРЕ в противном случае. Здесь V_0 - это некоторый пороговый уровень для минимального объема выборки. С точки зрения качества устной речи первостепенный интерес, безусловно, представляет собой множество четких ЭРЕ. Его, в таком случае, и следует считать основным итогом фонетического анализа речи.

Нетрудно понять, что именно в формулировке условий (2.2) и (2.3), а точнее, в выборе их порогов p_0 и V_0, кроются и основные возможности, и основные проблемы практического осуществления метода переменного дерева. Рассмотрим некоторые из них по результатам проведенных ниже экспериментальных исследований.

2.2. Результаты математического моделирования алгоритма фонетического анализа речи на основе адаптивной кластерной модели

Для проведения математического моделирования метода переменного дерева был выбран текст объемом около одной стандартной машинописной страницы, взятый из первой главы романа А.С. Пушкина "Капитанская дочка".

Этот текст был проговорен в среднем темпе группой дикторов. Полученные сигналы сначала записывались в память ПК в виде соответствующих звуковых файлов. Для этого применялась специальные программные и аппаратные средства: динамический микрофон AKG D77 S и ламповый микрофонный предусилитель ART TUBE MP Project Series USB. Частота дискретизации встроенного АЦП была установлена равной 8 кГц — общепринятая частота при обработке устной речи. Продолжительность записи по каждому диктору составила около полутора минут.

Далее для каждого диктора по алгоритму (1.9), (2.1)...(2.3) были выделены соответствующие множества ЭРЕ. При этом длина одного сегмента данных во всех случаях составляла m = 80 отсчетов, или 10 мс по времени. А для расчета коэффициентов авторегрессии в (1.9) применялась рекуррентная процедура Берга-Левинсона (1.16).

Порядок АР-модели (1.8) был установлен равным p = 20 в соответствии с результатами, полученными в работе [95]. Порог по величине информационного рассогласования в (2.2) для объединения в один класс множества реализаций каждой ЭРЕ варьировался в широких пределах p_0 = 0,5...3,0. Пороговый уровень в (2.3) для минимальной длительности четко проговоренных фонем варьировался в пределах V_0 = 240...800 отсчетов (30...100 *мс)*. Это два основных параметра разработанного выше алгоритма автоматического анализа речи (1.9), (2.1)... (2.4). В дальнейшем исследовалась зависимость от каждого из них фонетического состава речи разных дикторов.

На рис. 2.1 показана зависимость от значений p_0 и V_0 числа выявленных ЭРЕ для первого диктора. Видно, что при увеличении обоих порогов сначала происходит резкое уменьшение количества выделенных ЭРЕ. Проблема выбора их оптимальных значений решается здесь очевидным путем: p_0* =1,0... 1,2 и V_0*=3…7 — по принципу относительной стабилизации фонетического состава речевого сигнала. С одной стороны, при малых значениях каждого из порогов мы получаем чрезмерно большое количество ЭРЕ с небольшим различиями между собой в теоретико-информационном смысле. С другой стороны, при слишком

больших значениях порогов к одному классу ЭРЕ будут, возможно, относиться реализации заведомо разных ЭРЕ. А это безусловная ошибка фонетического анализа. Поэтому значения порогов V_0 и p_0 следует выбирать в тех точках на графике (рис. 2.1), где количество выделенных ЭРЕ достаточно представительно.

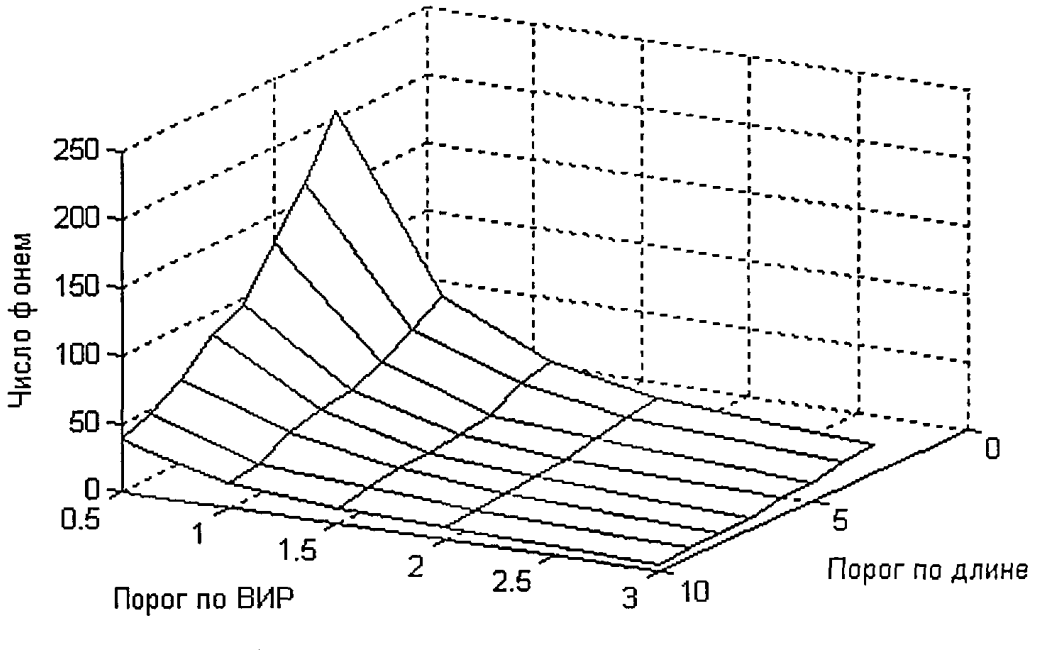


Рис. 2.1 - График числа выделенных фонем из речи первого диктора при различных порогах V_0 и p_0

Хорошей иллюстрацией к сказанному может служить временная диаграмма речевого сигнала на рис. 2.2, которая отображает короткий (2 с) фрагмент выбранного текста в произношении нашего диктора. Здесь для наглядности синхронно с текстом отмечены буквами алфавита соответствующие ЭРЕ русского языка. При этом параметры алгоритма обработки фиксировались постоянными: p_0= 1,1 и V_0 = 4. На этом рисунке в прямоугольники вдоль временной оси заключены сигналы для выявленных при автоматическом анализе четких ЭРЕ. Для всех них выполняется условие по минимальной длительности звучания (2.3). Остальные сигналы остались не идентифицированными и должны, поэтому рассматриваться как нечетко структурированная речь данного диктора. При

уменьшении значений обоих порогов количество и длина таких участков могут быть существенно сокращены, но только ценой понижения качества выделяемых из сигнала ЭРЕ.

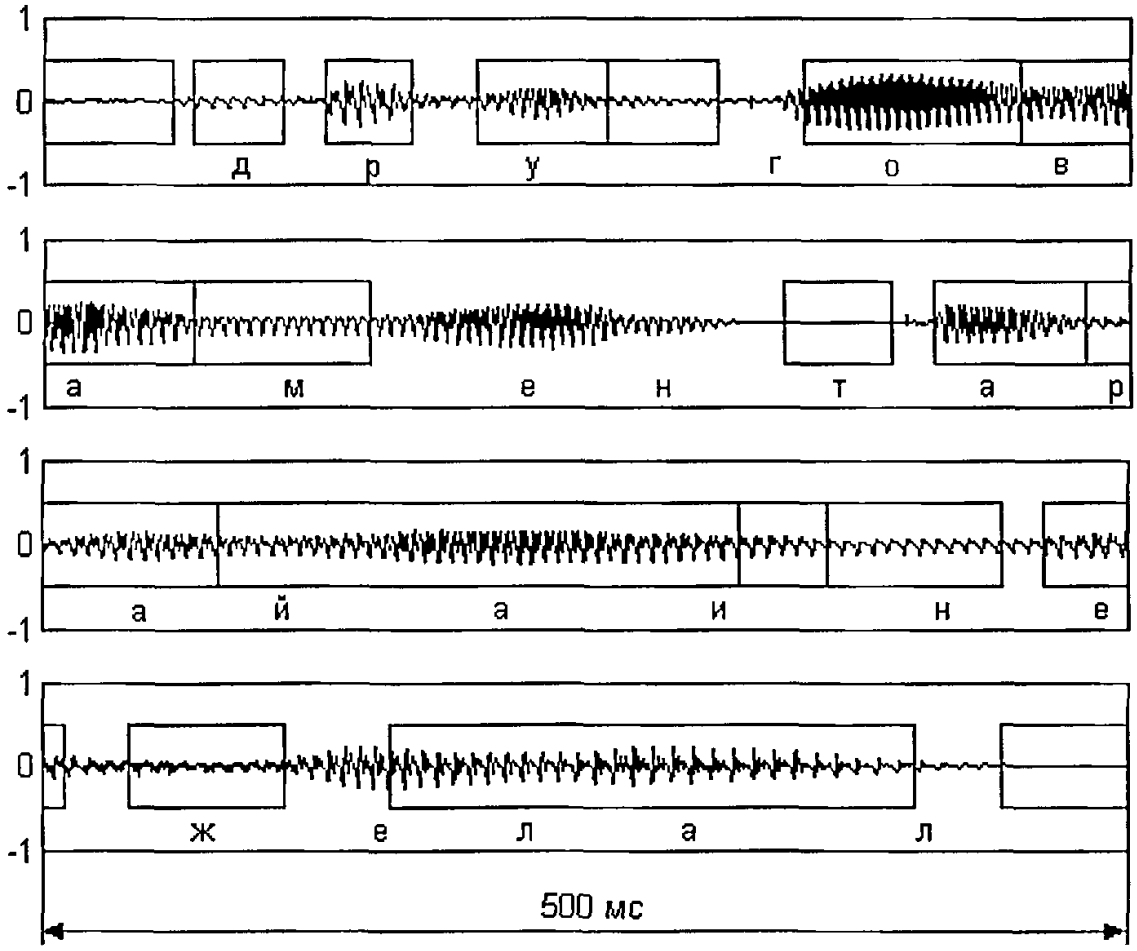


Рис. 2.2 - Участки с выделенными фонемами фразы «Другого ментора я и не желал»

Сделанные выводы остаются справедливыми для всех, без исключения, дикторов. В их подтверждение на рис. 2.3 представлена зависимость для второго диктора, аналогичная той, что изображена на рис. 2.1.

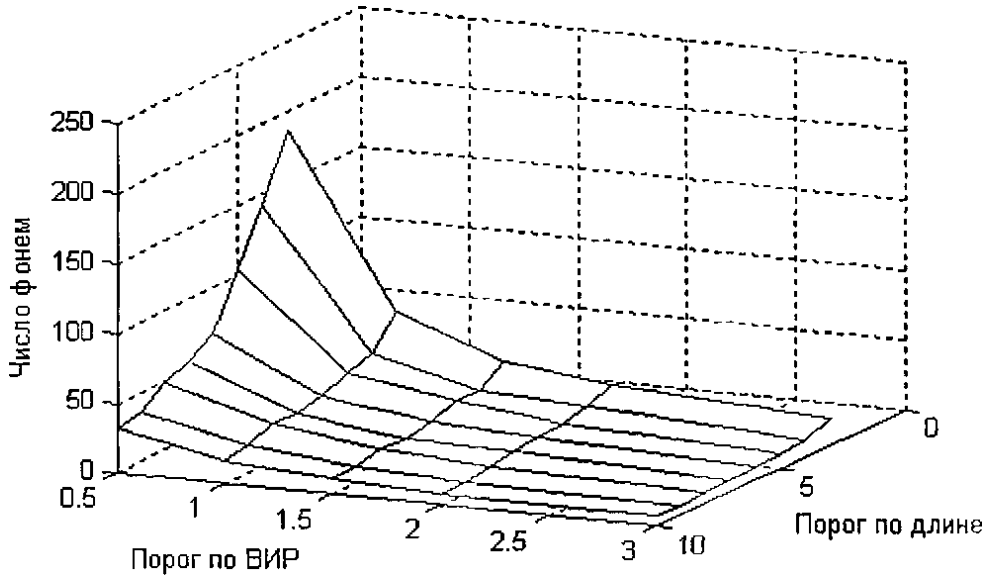


Рис. 2.3 - График числа выделенных фонем из речи первого диктора при различных порогах V_0 и p_0

Хорошо видно, что она качественно сохраняет прежний вид, хотя суммарное число выделенных ЭРЕ в этом случае оказалось существенно меньшим. Из сказанного можно сделать важный вывод об устойчивости оптимальных значений всех основных параметров обработки речи по методу переменного дерева (1.9), (2.1)...(2.4) по отношению к разным текстам, разным дикторам и их фонетическим особенностям речи.

2.3. Анализ эффективности алгоритма

Рассмотрим эффективность работы предложенного алгоритма. Для этого проведем аналитический и экспериментальный анализ вероятностей ошибок обнаружения фонем во входном сигнале.

Согласно работе [96] можно показать, что при гауссовой аппроксимации речевого сигнала вероятность ошибки первого рода (ложной тревоги)

$$\alpha = P\left\{x_M^2 > M\left(1 + p_0\right)\right\} \qquad (2.4)$$

определяется в первом приближении в терминах x^2_M – распределения Пирсона с M степенями свободы [97]. Значение M здесь ограничено сверху объемом T – 2Fτ выборки речевого сигнала X на интервалах τ его приблизительной стационарности, а также порядком p обеляющего фильтра в (1.9): $M \approx T - p$. Чем выше порог p_0, тем выше уровень значимости принимаемого согласно правилу (2.2) решения и тем более мягкими, следовательно, становятся требования в отношении качества сформированных диктором речевых единиц. И, наоборот, более жестким требованиям к качеству речи отвечает низкий порог в выражениях (2.2) и (2.4) и, соответственно, пониженный уровень значимости 1 - α.

С другой стороны, при заданном пороговом уровне p_0 вероятность ошибки второго рода: перепутывания v -й речевой единицы с r-й, равная

$$\beta_{rv} = P\left\{x^2_M < M(1+p_0)/(1+p_{rv})\right\}, r, v \le R, \qquad (2.5)$$

зависит, главным образом, от величины информационного рассогласования по Кульбаку-Лейблеру $p_{rv} \hat{=} p_r(x)\big|_{x=x_v}$ между рассматриваемой парой РЕ. Чем меньше значение p_{rv}, тем больше вероятность их перепутывания (2.5).

Для сравнения рассмотрим другой алгоритм с использованием известного критерия Пирсона, который также известен под названием «критерий x^2». Это непараметрический критерий для статистической проверки гипотезы о статистической связи между двумя переменными.

Проверка гипотезы о связи двух переменных осуществляется путем вычисления следующей статистики:

$$x^2_{выч} = \sum \frac{(P_i^{theor} - P_i^{emp})^2}{P_i^{theor}}, \qquad (2.6)$$

Где $P_i^{theor} = \int_{x_{i-1}}^{x_i} f(x)\,dx$ - предполагаемая вероятность попадания в i-й интервал,

$P_i^{emp} = \frac{n_i}{N}$ - соответствующее эмпирическое значение, n_i — число элементов выборки из i-го интервала. Полученная из (2.6) величина сравнивается с критическим значением, найденным по таблице распределения "хи-квадрат" для выбранного уровня значимости a и числа степеней свободы $L-1$, где L -

число наблюдений или число интервалов. Если $x^2_{выч} > x^2_{1-\alpha}$, нулевая гипотеза отклоняется в пользу альтернативной, согласно которой статистическая связь между переменными существует. Если $x^2_{выч} > x^2_{1-\alpha}$, оснований для отклонения нулевой гипотезы нет.

С учетом того, что речевой сигнал X хорошо аппроксимируется гауссовой случайной величиной [52], можно построить следующий алгоритм анализа фонетического состава речи:

Отсчеты анализируемого сигнала разбиваются на сегменты, по которым рассчитывается его эмпирическое распределение. Полученный результат сравнивается с теоретическим гауссовым распределением по критерию согласия Пирсона (2.6). В случае превышения статистикой значения $x^2_{1-\alpha}$ счетчик ошибок увеличивается на 1.

Для проведения экспериментальных исследований эффективности алгоритмов была выбрана пара фонем - «ш» и «ч», которые были проговорены диктором в режиме продолжительного, как наиболее информативного, звучания. Полученные сигналы с помощью стандартной аппаратуры и встроенного АЦП (частота дискретизации 8 кГц) записывались в память ПК в виде соответствующих звуковых файлов. Величина информационного рассогласования между фонемами составила $p_{v,r}$ = 0,726, а объёмами выборок - 10 и 12 тыс. отсчетов соответственно. С целью определения вероятности ошибки первого рода для разработанного алгоритма на обеляющий фильтр, настроенный на первую фонему, подавались короткие сегменты длиной Ь этой же фонемы. Отклик обеляющего фильтра сравнивался с пороговым уровнем $p_0 = 0,6$, и в случае его превышения счетчик ложных срабатываний $w_{ош}$ увеличивался на единицу. После обработки всех сегментов выбранной фонемы рассчитывалась величина ошибки первого рода α(L), как отношение $w_{ош}$ к общему числу сегментов w при заданном значении объема выборки: $\alpha(L) = \frac{w_{ош}(L)}{w(L)}$, где значение длины сегмента L изменялось от 30 до 300 отсчетов с шагом 10. Полученная зависимость (кривая 2 на рис. 2.4) была дополнена соответствующей

теоретической зависимостью вероятности ошибки первого рода от объема выборки (кривая 1 на рис. 2.4) и аналогичной экспериментальной зависимостью для алгоритма на основе критерия Пирсона (кривая 3 на рис. 2.4).

Из рис. 2.4 видно, что в целом экспериментальная зависимость хорошо согласуется с теоретической, что является дополнительным обоснованием в использовании метода обеляющего фильтра и гауссовой модели речевых сигналов в задаче тестирования качества устной речи. Динамические свойства разработанного алгоритма и алгоритма на основе критерия Пирсона могут быть охарактеризованы по средствам сравнения объемов выборочных данных, при которых достигается требуемый уровень вероятности ошибки первого рода. Так, вероятность ошибки, равная 5%, достигается для алгоритма периодизации на основе кластерной модели элементарных речевых единиц при объеме выборки N_1= 100, а для алгоритма на основе критерия Пирсона при N_2 = 270. Таким образом, скорость сходимости разработанного алгоритма существенно выше, чем у его аналога.

Для сравнения эффективности рассматриваемых алгоритмов были построены аналогичные зависимости для вероятности ошибки второго рода (рис. 2.5). При этом для построения экспериментальных зависимостей использовалась та же методика эксперимента, но на вход обеляющего фильтра подавались короткие сегменты не первой, а второй фонемы - «ч». Тогда вероятность ошибки второго рода $\beta(L)$ определялось как отношение числа пропусков разладки $w_{ош}$ (не превышение откликом обеляющего фильтра порога допустимой разладки при периодизации) при обработке w сегментов: $\beta(L)=\frac{w_{ош}(L)}{w(L)}$. Из графика видно, что при заданной вероятности

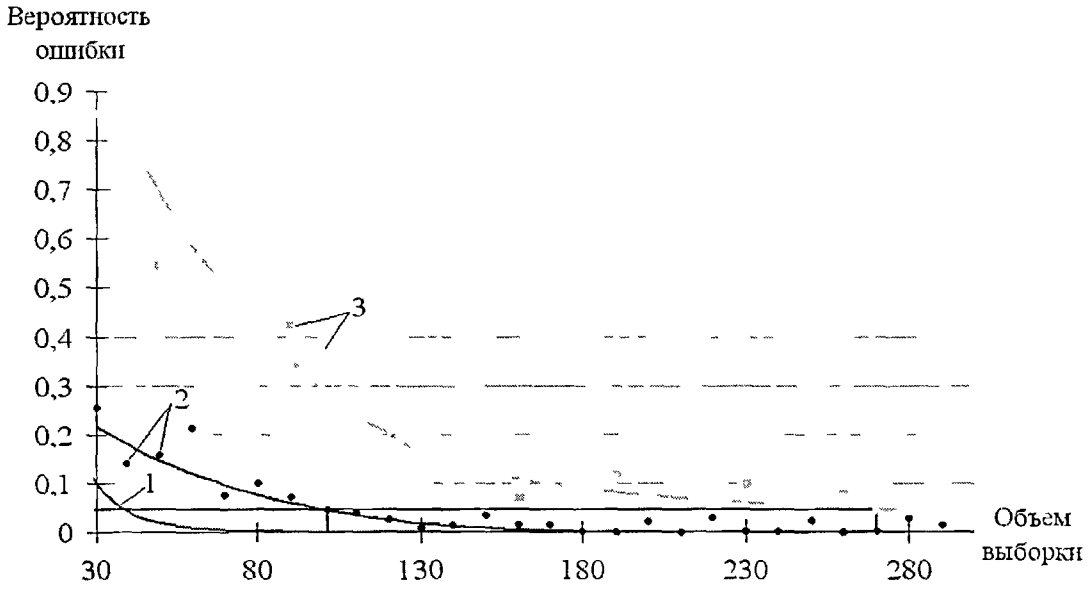


Рис. 2.4. Зависимость вероятности ошибки первого рода от объёма выборки в задаче анализа эффективности алгоритмов:

1 — теоретическая зависимость для алгоритма на основе кластерной модели элементарных речевых единиц;

2 — экспериментальная зависимость для алгоритма на основе кластерной модели элементарных речевых единиц;

3 - экспериментальная зависимость для алгоритма на основе критерия согласия Пирсона.

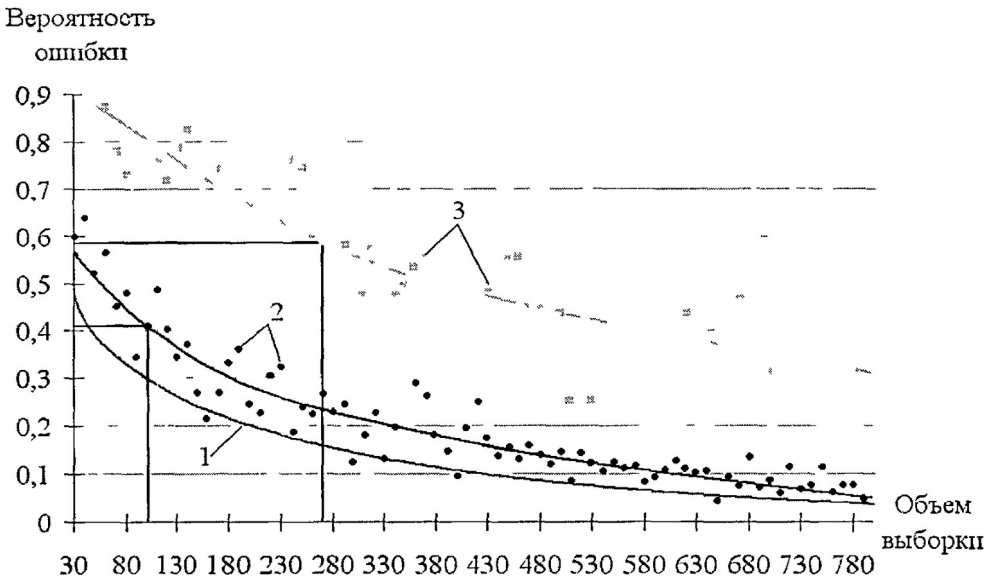


Рис. 2.5. Зависимость вероятности ошибки второго рода от объёма выборки в задаче анализа эффективности алгоритмов:

1 - теоретическая зависимость для алгоритма на основе кластерной модели элементарных речевых единиц;

2 - экспериментальная зависимость для алгоритма на основе кластерной модели элементарных речевых единиц;

3 - экспериментальная зависимость для алгоритма на основе критерия согласия Пирсона.

ошибки первого рода (при N_1=100 и N_2 =270) вероятность ошибки второго рода для алгоритма на основе кластерной модели элементарных речевых единиц составляет β_1 = 42%, а для алгоритма на основе критерия Пирсона - β_2 =59%. Таким образом, эффективность предложенного алгоритма в 1,4 раза выше по сравнению с аналогом. Указанный выигрыш можно объяснить не выполнением на практике в большинстве случаев требования независимости выборки, которое предъявляется к критерию согласия Пирсона.

Следует отметить, что абсолютные значения вероятностей ошибок второго рода были получены для проблемной пары реализаций фонем, т.е. для случая, когда $p_{v,r} \approx p_0$. Именно для таких фонем возникает большинство проблем при распознавании. Величины информационного рассогласования для большинства фонем значительно превосходят пороговый уровень $p_{v,r} \approx p_0$, поэтому соответствующие вероятности ошибок второго рода для них малы.

Практическая реализация предложенного алгоритма фонетического анализа слитной речи (1.9), (2.1)...(2.4) была положена в основу создания информационной системы фонетического анализа слитной речи [99], краткое описание и основные возможности которой приведены в следующей главе.

Выводы по главе

В главе предложен алгоритм формирования адаптивной кластерной модели элементарных речевых единиц для создания новых и обновления существующих ФБД. Его ключевым звеном является метод (R+1)-элемента и критерий минимального информационного рассогласования. Проведены

экспериментальные исследования алгоритма, в результате которых были сделаны следующие выводы: пороговые значения V_0* и p_0* должны устанавливаться в диапазонах 3...7 и 1,0...1,2 соответственно. Причем данные значения остаются справедливыми для всех без исключения дикторов.

Разработанный алгоритм является мощным инструментом для фонетического анализа слитной речи: варьируя значения порогов по величине информационного рассогласования и длине ЭРЕ можно в широких пределах изменять его чувствительность.

Анализ эффективности работы алгоритма показал, что он позволяет достичь, по крайней мере, в 1,4 раза меньшей вероятности перепутывания фонем, чем, например, метод на основе известного критерия Пирсона. Следует отметить, что такой анализ проводился по паре проблемных фонем, для которых возникает наибольшее количество ошибок.

ГЛАВА 3. Результаты экспериментальных исследований алгоритма фонетического анализа речи на основе адаптивной кластерной модели

3.1. Лабораторный образец фонетического анализатора речи

Для реализации предложенного алгоритма фонетического анализа слитной речи был разработан лабораторный образец информационной системы фонетического анализа слитной речи (ИС ФАР). Данная система представляет собой фонетический анализатор. Варианты применения такого анализатора можно привести из самых различных областей. Это может быть, например, задача анализа качества речи по ее фонетическому составу, как для отдельного диктора, так и для группы дикторов, народности или даже всего национального языка. В качестве прикладной задачи можно привести сжатие речевых данных и их передачу по каналам связи при сохранении высокого качества речи.

ИС обеспечивает выполнение следующих возможностей:

1. автоматическое выделение фонем из входных данных;
2. обработка фонем;
3. хранение фонем;
4. визуализация полученных результатов.

Блок-схема ИС ФАР показана на рис.3.1.

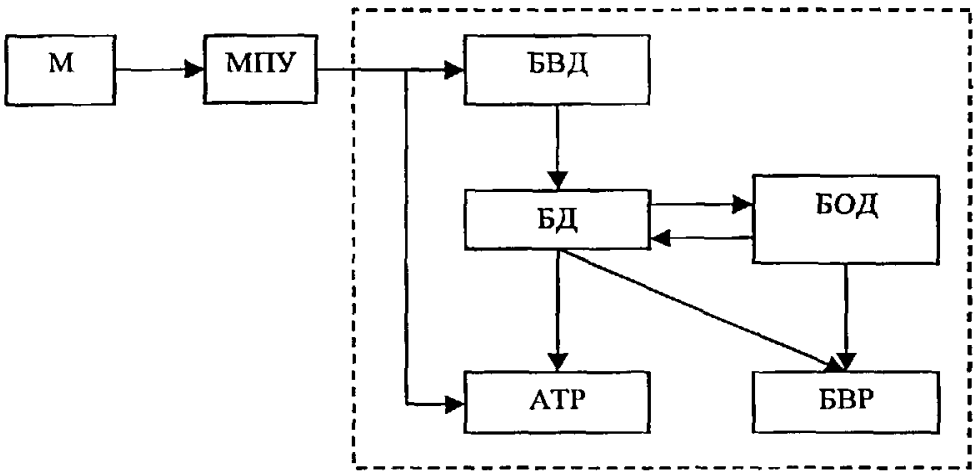


Рис. 3.1 - Блок-схема ИС ФАР

Здесь М — динамический микрофон AQG D77 S. МПУ — микрофонный ламповый предусилитель со встроенным аналого-цифровым преобразователем. БВД - блок ввода данных. БВР — блок вывода результатов, БД — база данных, БОД — блок обработки данных, АТР - подсистема автоматического транскрибирования речевых сигналов. Блоки, ограниченные штриховой линией, выполнены в виде соответствующих программных модулей и составляют собственно информационную систему. Звуковой сигнал поступает с микрофона М на предусилитель МПУ, который осуществляет его усиление и преобразование в цифровой вид. Преобразованный таким образом сигнал поступает по шине USB в персональный компьютер, где осуществляется его запись в звуковой файл. Этот файл считывается БВД, где осуществляется его предварительная обработка, анализ и запись полученных результатов в базу данных в соответствии с выбранным режимом обработки.

БВР осуществляет извлечение данных из БД, соответствующих требуемому критерию, и отображение их в виде, удобном для пользователя. БОД предназначен для организации возможности работы с группами дикторов и формирования входных данных для БВР.

Подсистема АТР также позволяет производить отображение в различных режимах и автоматическую разметку входного сигнала на фонемы в соответствии со списком фонем какого-либо диктора из БД. Кроме того, возможно озвучивание выбранных фрагментов сигнала, сохранение результатов транскрибирования в текстовый файл и анализ получаемых результатов.

Информация, содержащаяся в БД, может быть, как непосредственно считана БВР, так и может поступать в БВР через БОД.

Разработанная информационная система использует базу данных формата Microsoft Access. На рис. 3.2 показана схема этой базы данных.

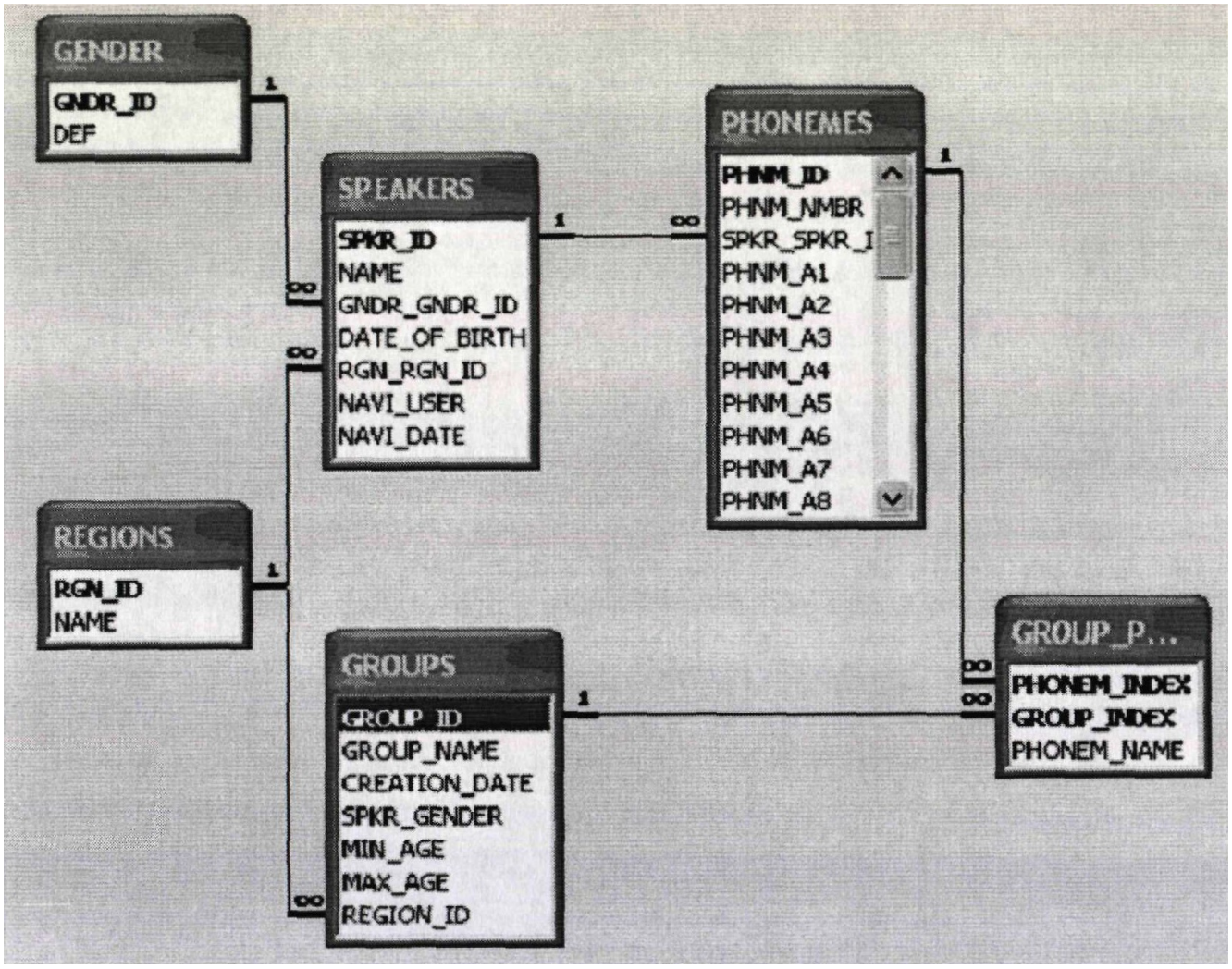


Рис. 3.2 - Схема базы данных ИС ФАР

БД включает в себя следующие таблицы:

- GENDER - таблица-справочник содержит значения, определяющие пол диктора: мужской или женский.
- SPEAKERS - таблица хранит информацию о дикторе: имя диктора, его пол, год рождения, регион, в котором проживает данный диктор.
- » REGIONS - таблица-справочник хранит информацию о названиях регионов.
- PHONEMES - таблица хранит АР-коэффициенты и названия фонем дикторов.
- » GROUPS - таблица хранит информацию о группах дикторов: название группы, дату создания, пол дикторов, входящих в группу, максимальный и минимальный возраст дикторов, регион.

• GROUPPHONEMES - таблица служит для организации связи "многие-ко-многим" между таблицами PHONEMES и GROUPS; содержит наименование фонем данной группы, определяемые пользователем на слух.

Интерфейс ИС ФАР состоит из главной формы, на которой отображаются дикторы, внесенные в базу данных и главное меню программы. При выборе любого диктора из списка в правой части окна отображается краткая информация о нем. Кроме того, при помощи имеющегося меню можно выбирать различные режимы работы, загрузки, сохранения и отображения данных. Общий вид интерфейса показан на рис. 3.3.

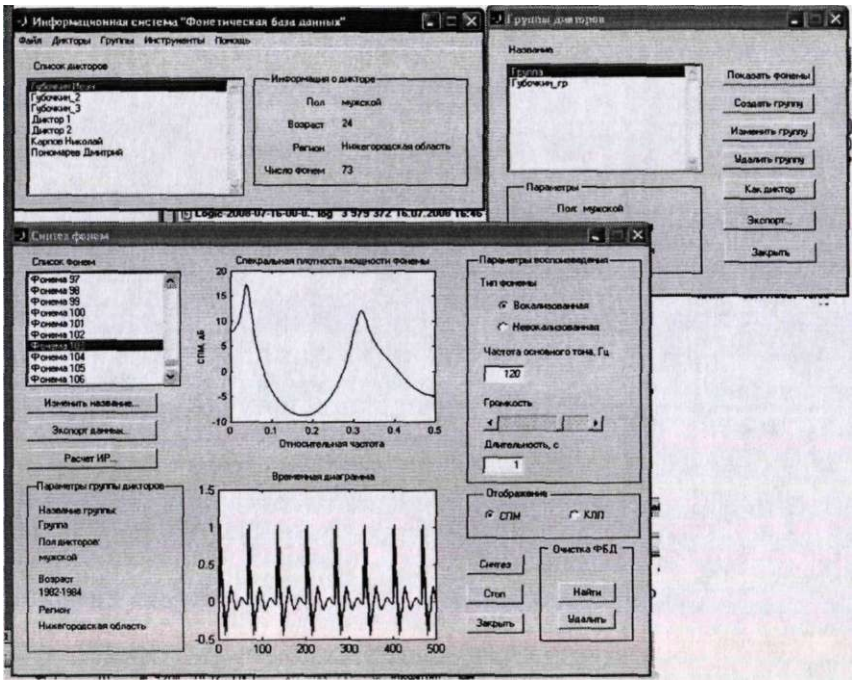


Рис. 3.3 - Вид пользовательского интерфейса

На главной форме отображаются дикторы, внесенные в базу данных и главное меню программы. При выборе любого диктора из списка в правой части окна отображается краткая информация о нем.

Ниже приводится краткое описание каждого из разделов главного меню.

- Меню "Файл" содержит единственное подменю "Выход", которое служит для выхода из программы.
- Меню "Дикторы" содержит функции для работы с дикторами и фонемами дикторов:

- "Добавить" — позволяет добавить нового диктора в БД.
- "Изменить" — позволяет изменить имя и информацию о дикторе,
- "Удалить" - удаляет выбранного диктора,
- "Удалить все" — удаляет всех дикторов из списка,
- "Фонемы" - содержит функции для работы с фонемами выбранного диктора.

- Меню "Группы" содержит функции для организации работы с группами дикторов:
  - "Показать" - открывает форму для работы с группами дикторов.
- Меню "Инструменты" содержит функции для расчета ВИР и настройки системы.
  - "Расчет ВИР" — открывает форму для расчета ВИР между фонемами разных дикторов или групп дикторов,
  - "Настройки" — открывает форму настроек системы.
- "Помощь" - содержит функцию вывода справки.
  - "О программе" - выводит диалог с краткими сведениями о разработанной программе.

На рис. 3.4 показана форма настроек информационной системы. Здесь задаются основные параметры для работы реализованных в ней алгоритмов.

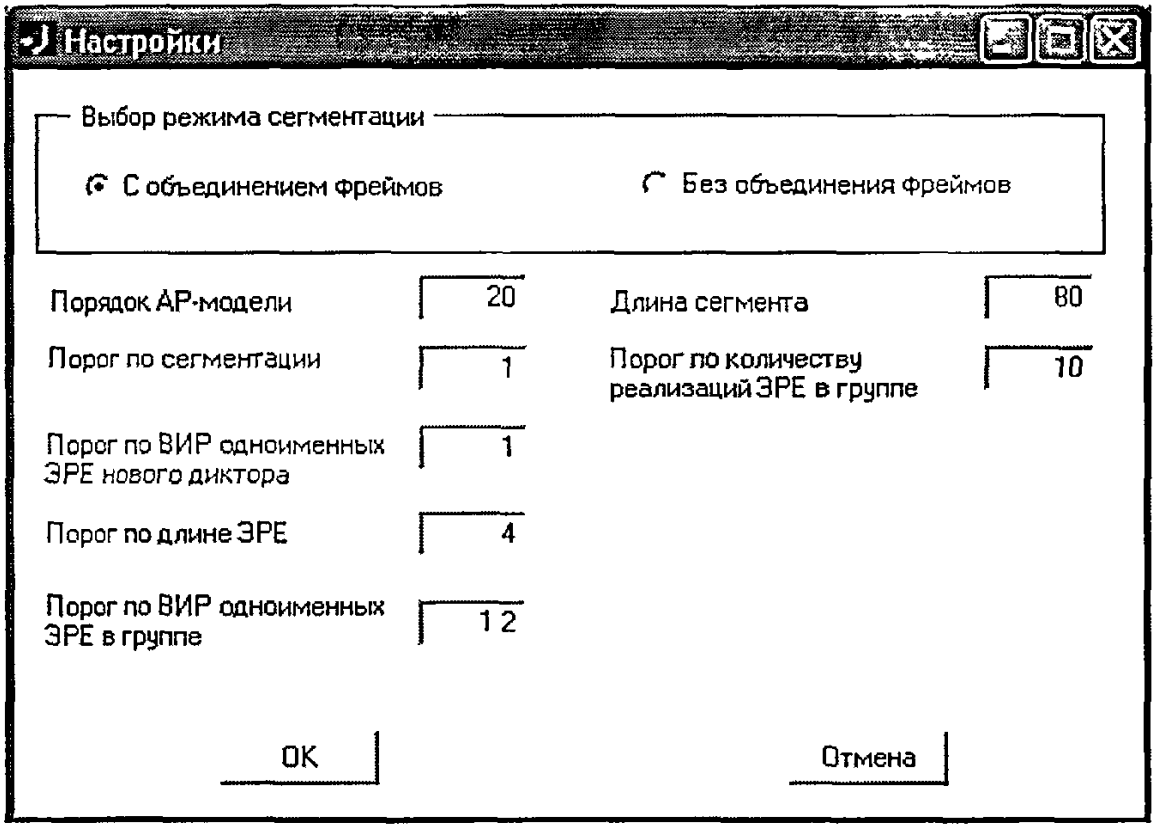


Рис. 3.4 - Окно настроек

Порядок АР-модели — целое число, большее единицы. Рекомендуется задавать в пределах 20...30. Порог по сегментации — любое число больше 0. Рекомендуется задавать в диапазоне от 0,8 до 1,5 (порог разладки при сегментировании должен быть больше 0.5). Этот порог используется на этапе сегментирования входного сигнала на фонемы. Порог по величине информационного рассогласования одноименных ЭРЕ нового диктора - любое число больше 0. Рекомендуется задавать в диапазоне от 0,8 до 2,0. Этот порог используется при объединении фонем, выделенных на этапе сегментирования в классы. Порог по длине ЭРЕ - целое число больше 0. Рекомендуется задавать в интервале от 3 до 7. Этот порог используется для задания минимального количества сегментов, которые должна включать в себя фонема, для того чтобы она могла участвовать в процедуре классификации. Порог по величине информационного рассогласования одноименных ЭРЕ в группе - любое число больше 0. Рекомендуется задавать в интервале от 0,8 до 2,0. Этот порог используется при наполнении фонемами группы дикторов.

Длина сегмента задается в отсчетах и по умолчанию равна 80. Рекомендуется задавать его в пределах от 80 до 320. Порог по количеству реализаций ЭРЕ в группе определяет минимальное количество выделенных реализаций, относящихся к одной фонеме, при котором данная фонема будет включена в базу данных. По умолчанию значение данного порога равно 10.

На рис. 3.5 показана форма расчета величины информационного рассогласования между фонемами разных дикторов или групп. Соответствующий режим выбирается кнопками "По дикторам" и "По группам".

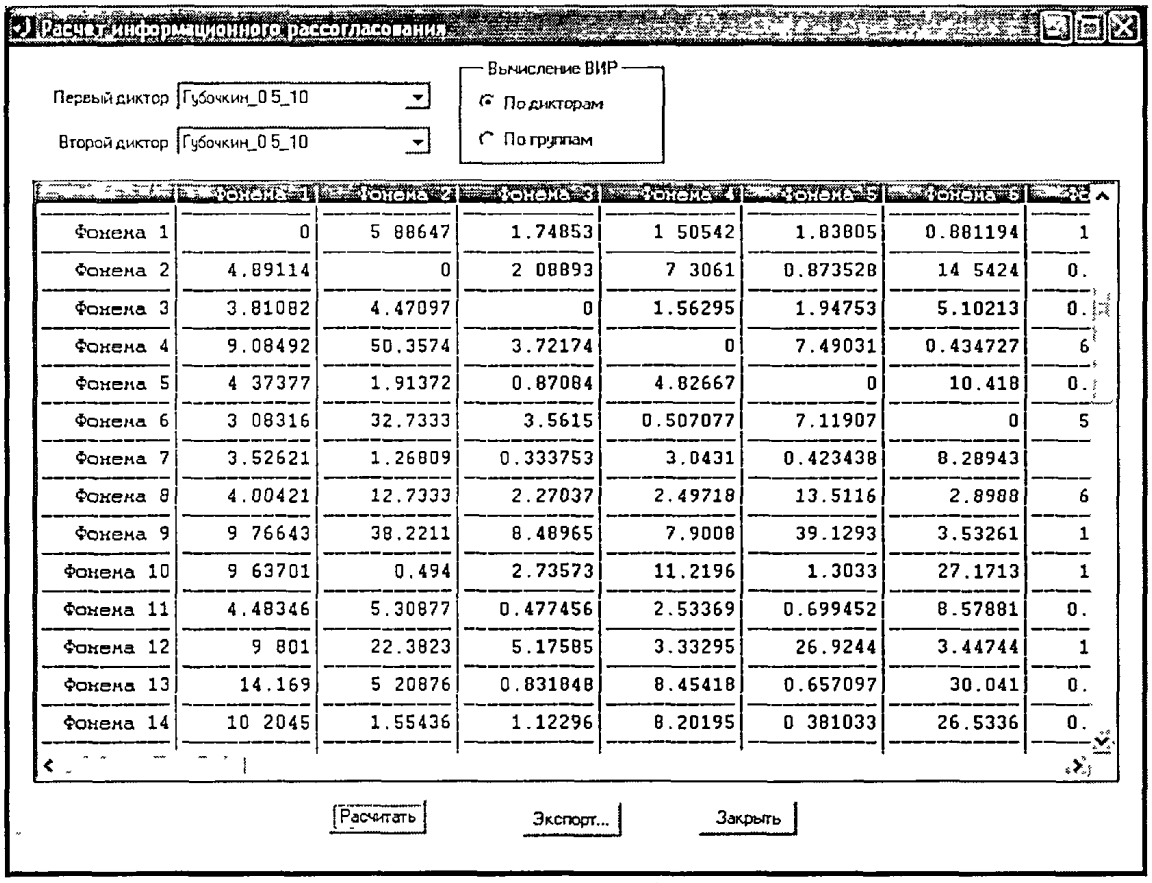


Рис. 3.5 - Форма расчета ВИР между фонемами разных дикторов или групп

3.2. Повышение точности алгоритма распознавания речи

Повышение точности распознавания речи на сегодняшний день является одной из ключевых задач АРР. В рамках теоретико-информационного подхода на основе кластерной модели ЭРЕ ее решение заключается в учете того факта, что различные сегменты анализируемого слова X могут вносить различный вклад в сумму (1.14) предложенного выше алгоритма (1.10)...(1.17). Часть сегментов выполняет служебные функции и с точки зрения задачи АРР может не учитываться. Такая идея реализуется, в частности, в методе обеляющего фильтра с амплитудным ограничением сверху парциальных значений p_{1r}(X,k) решающей статистики (1.10) [83]:

$$p_{1,r}(X,k)=\begin{cases} U, & при\ \ \sigma^2_{1,r}(X,k)-1>U, \\ \sigma^2_{1,r}(X,k)-1, & при\ \ \sigma^2_{1,r}(X,k)-1\le U \end{cases},\ k=\overline{1,N_r},\ r=\overline{1,R},\ l=\overline{1,L}. \qquad (3.1)$$

Здесь $\sigma^2_{1,r}(X,k)$ - выборочные дисперсии откликов ОФ P-го порядка, настроенного на k-й участок r-го эталонного слова и на l-й сегмент сигнала на входе, U— пороговое значение.

Экспериментальные исследования модифицированного алгоритма АРР проводились на речевой базе, описанной в разделе 1.3. Целью исследований являлось сравнение алгоритма АРР (1.10)...(1.17) и модифицированного алгоритма (1.10)...(1.17), (3.1) с амплитудным ограничением величины решающей статистики, а также исследование влияния порога амплитудного ограничения $\boldsymbol{U}$ на вероятность правильного распознавания слов. Результаты экспериментальных исследований иллюстрируются графиками и таблицами ниже по тексту.

В таблице 3.6 показаны оценки вероятности правильного распознавания десяти слов из подготовленной экспериментальной базы для двух различных алгоритмов распознавания: по сегментам с решающей статистикой (3.1) и по сегментам с решающей статистикой (1.10). Оценка вероятности правильного

распознавания слов вычислялась по результатам серии испытаний и формуле относительной частоты случайных событий.

Таблица 3.6 - Вероятности правильного распознавания слов для решающей статистики без ограничения (1.10) и с ограничением (3.1)

Слово	Алгоритм	
	Для решающей статистики (1.10)	Для решающей статистики (3.1) при пороге U=0,5
0	0,89	1
1	0,98	1
2	0,99	1
3	0,97	0,98
4	0,99	1
5	0,95	0,99
6	0,97	0,97
7	0,92	0,98
8	0,95	1
9	0,96	0,98
Итого	**0,957**	**0,99**

Представленные результаты указывают на существенное уменьшение ошибок перепутывания слов, за счет амплитудного ограничения слагаемых в решающей статистике (3.1).

Повышение точности распознавания при использовании амплитудного ограничения слагаемых в решающей статистике (3.1) можно объяснить тем, что в анализируемом слове X могут появляться сегменты, имеющие большое значение ИР и тем самым существенно увеличивающие значение результирующей решающей статистики (1.14). Подобные явления могут рассматриваться как «импульсные помехи» и для их подавления необходимо

использовать методы, аналогичные тем, что широко применяются в статистической радиотехнике [84]. Применение амплитудного ограничения с порогом и в выражении (3.1) позволяет эффективно бороться с импульсными помехами и накапливать к моменту принятия решения лишь информативные составляющие решающих статистик. Причём, в зависимости от значения порога и будем получать различные ошибки перепутывания слов.

В таблице 3.7 показаны вероятности правильного распознавания слов для решающей статистики (3.3) в зависимости от значения порога U. Видно, что максимальная вероятность правильного распознавания достигается при U=0,5. Таким образом, данное значение ограничения можно выбрать в качестве оптимального.

Таблица 3.7 — Зависимость вероятности правильного распознавания слов от величины порога U

Слово	**Порог**						
	0,3	**0,5**	**1**	**1.5**	**2**	**2,5**	**3**
0	0,99	**1**	**1**	**1**	0,99	0,98	0,98
1	**1**	**1**	**1**	**1**	**1**	**1**	**1**
2	**1**	**1**	**1**	**1**	0,98	0,97	0,96
3	0,98	0,98	0,98	0,98	0,98	0,97	0,97
4	0,99	**1**	0,99	0,99	0,99	0,99	0,99
5	0,96	0,99	0,98	0,98	0,97	0,95	0,95
6	0,96	0,97	0,98	0,98	0,99	0,99	0,99
7	0,82	0,98	0,87	0,72	0,67	0,66	0,66
8	**1**	**1**	**1**	**1**	**1**	**1**	**1**
9	0,98	0,98	0,99	0,99	0,99	0,99	0,98
Итого:	**0,968**	**0,99**	**0,979**	**0,964**	**0,956**	**0,95**	**0,948**

Повышение точности АРР при использовании статистики (3.1) можно объяснить, рассматривая эффект накопления в ней величины ИР из выражения (1.13). Зависимость суммарного информационного рассогласования (1.14)

между сигналом X и истинным словом-эталоном от количества анализируемых сегментов при применении двух статистик (1.10) (сплошная линия) и (3.1) (пунктирная линия) для оптимального порога U=0,5 показана на рис. 3.6 для одной из реализаций слова "ноль".

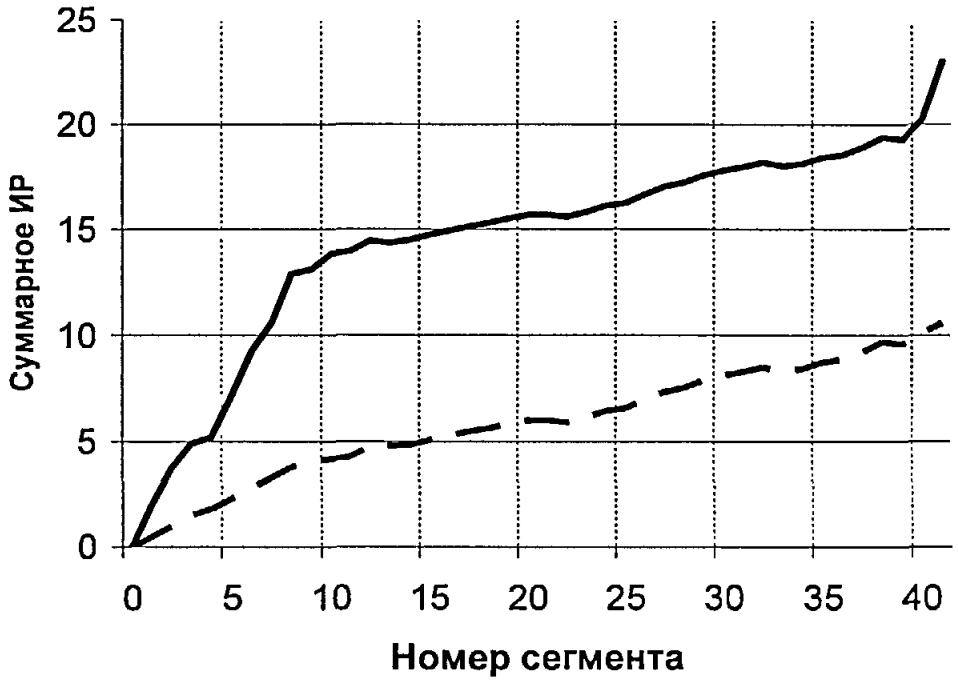


Рис. 3.6 - Сравнение зависимости накопления суммарного ИР при применении двух статистик (1.10) (сплошная линия) и (3.3) (пунктирная линия)

Из графиков видно, что информационное рассогласование для статистики (1.10) резко возрастает на ряде сегментов анализируемого слова. Использование же амплитудного ограничения в решающей статистике (3.1) позволяет учитывать только информативные сегменты, что повышает надежность разработанного алгоритма автоматического распознавания речи.

В предложенном алгоритме распознавания речи могут также применяться и другие пути повышения быстродействия. Наиболее перспективным подходом в данном случае можно считать использование метода направленного перебора словаря и его модификаций [87, 115].

Выводы по главе

Приведены результаты практического применения рассмотренных выше методов в рамках разработанной информационной системы фонетического анализа слитной речи.

Данная информационная система обладает уникальным спектром своих функциональных возможностей, вытекающих из новых возможностей информационной теории восприятия речи, и гибким интерфейсом, позволяющим в полной мере раскрыть данные возможности. ИС может быть использована при решении множества актуальнейших проблем в области фонетического анализа речи: и вариативности устной речи, и априорной неопределенности, и сжатия данных, и вычислительной сложности, и, главное, проблемы обновления речевых баз данных без разрушения их структуры.

Также в главе предложен информационный критерий сравнения ПФБД двух дикторов. Показана возможность его использования при выборе оптимальной ПФБД из множества. Сделан вывод о том, что для надежного нахождения оптимальной фонетической кластерной базы данных внутри группы достаточно 10 — 15 дикторов.

Предложен метод повышения точности распознавания речи учитывающий существенно неравноценный вклад различных сегментов анализируемого слова в величину итоговой решающей статистики. Показано, что использование амплитудного ограничения на величину решающей статистики для отдельных сегментов позволяет повысить вероятность правильного распознавания слов более чем на 4% (с 0,95 до 0,99 в проведенном эксперименте).

ГЛАВА 4. Применение адаптивной кластерной модели в новой технологии информационного обеспечения задачи обучения речи слабослышащих

4.1. Обзор существующих подходов к решению задачи обучения речи с визуализацией речевого сигнала

В настоящее время проблема обучения речи людей с нарушениями слуха является весьма актуальной. Общеизвестно, что страдающие нарушениями слуха не могут четко и правильно произносить слова и даже отдельные звуки, поскольку не слышат того, что говорят, и потому не могут контролировать свое произношение. Современная дефектология имеет широкий арсенал методов для постановки произношения у глухих и слабослышащих детей [101 — 103]. Однако все они требуют постоянного вмешательства логопеда и практически не предполагают индивидуальной работы ребенка над собой. Вместе с тем, для сохранения разборчивости речи слабослышащим, а особенно глухим требуется постоянная тренировка собственного произношения. Значительную помощь в данной ситуации могут оказать автоматизированные системы контроля.

Компьютеры и другие технические устройства предоставляют большое число возможностей по улучшению произношения через визуализацию артикуляционного процесса при помощи различных средств отображения [104]. Здесь применяется большое число методов и подходов. Рассмотрим некоторые из них более подробно.

Системы отображения речевого тракта. В данных системах сделано множество попыток для как можно более точного отображения состояния голосового тракта говорящего, целью которых была помощь в обучении произношению отдельных гласных. Две из таких систем - «Computer Vowel Trainer» [105] и «Vocal Tract Area Display» [29] используют коэффициенты

линейного предсказания (КЛП) для приближенного отображения формы речевого тракта.

IBM SpeechViewer. Эта система является одной из самых лучших на сегодняшний день. Она была выпущена как коммерческий продукт в 1989 году [30]. IBM SpeechViewer - это акустическая система, которая позволяет получать в реальном времени визуальную обратную связь. В России данная система известна под названием «Видимая речь».

Indiana Speech Training Aid (ISTRA). Система ISTRA [31 - 33] работает на принципах, схожих с IBM SpeechViewer, но имеет плату распознавания изолированных слов с настройкой на диктора. Система обрабатывает звуковой сигнал и выводит численное значение степени его близости к некоторому эталону. Главное отличие от IBM SpeechViewer состоит в том, что ISTRA лучше приспособлена к распознаванию отдельных слов.

The Video Voice Speech Training System. Это коммерческая система, производимая компанией Microvideo, позволяет анализировать звуковой сигнал и выделять из него различные параметры, такие как громкость, частота основного тона и спектр [34]. Здесь параметры речи пользователя также сравниваются с некоторыми эталонными значениями.

Особенностью данной системы является наличие окна с отображением формантных частот «F1 vs. F2», который помогает корректировать произношение гласных и артикуляцию в пределах отдельных слов.

Speech Rehabilitation Speech Training Aid for the Hearing Impaired (HARP). HARP [35 - 37] использует похожие на Video Voice экраны для обучения произношению звуков в реальном времени. Для анализа звуков используется метод крос-дикторной нормализации, позволяющий получить двумерное представление качества звука по двум первым формантам и частоте основного тона.

Среди отечественных работ по созданию систем обучения речи следует отметить программу «Контроль произношения» [38]. Эта система оценивает

качество произнесенной учеником фонемы путем ее сравнения с набором эталонов.

Среди отечественных коммерческих систем наиболее известны следующие: "Речевой калейдоскоп" - НПО "КАСКОД" — С.Петербург; "Филиппок" - КУДИЦ - Москва; "КОСПР" — Минск. Практически все они использовали нестандартные внешние устройства обработки речи, заменяющее современные звуковые карты, и являлись аналогами разработок IBM. В настоящее время из отечественных разработок предлагаются к поставке тренажеры "Дельфа" - Москва, "Учимся говорить" - С.Петербург, "Визуальный тренажер произношения" — Минск.

Главный принцип, реализуемый в рассмотренных системах постановки произношения - это сравнение исходного сигнала с некоторым эталоном. Основной недостаток такого подхода заключается в том, что диктор (обучаемый) в силу ряда причин, может быть в принципе не способным точно повторить эталон. Для его преодоления проф. В.В. Савченко была разработана информационная теория обучения речи (ИТОР), в основе которой лежит критерий минимального информационного рассогласования в метрике Кульбака-Лейблера. Основные положения данной теории приведены в следующем разделе.

4.2. Визуализация речевого сигнала на основе кластерной модели речевых единиц

В работе [106] предлагается сравнение тестируемого сигнала x одновременно с несколькими эталонами $\boldsymbol{x}_{r,j}$, $j=\overline{1,J_r}$, по каждой r-й РЕ $(r=\overline{1,R})$. Диктору будет достаточно приблизить свое произношение к любому из них: каждый конкретный диктор в процессе своего обучения выбирает наиболее удобный, достижимый для себя вариант эталонного произношения из заданного множества альтернатив $x_{r,j}$. Рассмотрим данный
подход более подробно.

Согласно ИТОР, можно утверждать, что одноименные реализации РЕ в сознании человека группируются в соответствующие классы вокруг некоторого центра - эталона данного образа. Отсюда становится очевидным и механизм постановки произношения по каждой отдельной РЕ изучаемого языка: тестируемый сигнал x в процессе обучения конкретного диктора должен войти в состав множества допустимых решений X_r по критерию

$$p_r^*(x) = p\left(\frac{x}{x_{rv}}\right) \le p_0(1\ldots2)\, p_r^* \qquad (4.1)$$

ограничения сверху на величину ИР относительно соответствующего ИЦ-эталона x_r^* из выражения (0.1).

Отобразим отдельные реализации данной РЕ в виде точек на плоскости (рис. 0.1). Тогда точка с минимальной суммой расстояний выступит в роли своеобразного "центра массы" заданного множества (отмечена на рис. 0.1 залитым элементом). В этом и состоит физический смысл ИЦ-эталона (0.1). Тогда множество рассматриваемых реализаций данной РЕ — это допустимые вариации эталона с точки зрения его теоретико-информационных свойств, а пороговый уровень p_0 из выражения (4.1) - это количественные требования к допустимому уровню упомянутых вариаций. В данном механизме реализуется указанный ранее принцип МИР [49, 71, 107].

Наглядная геометрическая интерпретация процесса обучения речи (0.1), (4.1) представлена на диаграмме (рис. 4.1). Здесь точками на плоскости обозначено имеющееся множество эталонов рассматриваемой РЕ, а каждый кружок - это тестируемый сигнал на очередном шаге обучения диктора. "Траектория обучения" отображается направленной ломаной линией. Процесс обучения сводится к итеративному (пошаговому) приближению каждого очередного варианта сигнала **x** к центру "области допустимых решений", обозначенной на рисунке серым фоном. Границу области образует окружность радиусом p_0 относительно ИЦ-эталона x_r^*.

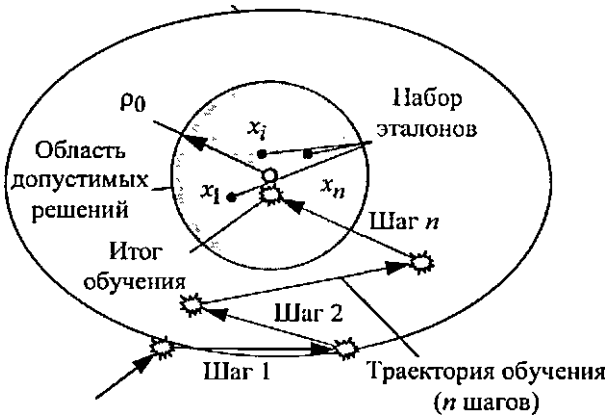


Рис. 4.1 - Схема обучения на основе ИТОР

Основное значение здесь имеет привязка порога p_0 к минимальной средней величине p_r* информационного рассогласования $p(x_{rj}/x_{rv>})$ на множестве выборочных образцов по каждой отдельной РЕ $X_r = \{x_{rj}\}$ конечного объема $J<\infty$. Регулируя величину порога p_0, мы регулируем требования к качеству речи обучаемого. Например, для дикторов с дефектами слуха или речеобразования эти требования могут быть очень низкими (p_0=max). При обучении же профессиональных дикторов они наоборот повышаются до максимума (p_0 = min). Кроме того, может меняться и используемая база эталонов $\{x_r^*\}$. В одних случаях это образцы высококачественной речи, а в других - речевой материал, несущий в себе типичные черты говора жителей определенной местности, возраста, пола и т.п. Реализация описанного алгоритма при помощи метода «улитки» приведена в работе [108].

На основе технологии обучения, описанной выше, был разработан лабораторный образец информационной системы для обучения речи глухонемых и слабослышащих (ИСОР) [109 - 113]. Данная система может функционировать в двух режимах: фонетическом и морфологическом. При работе в фонетическом режиме проводится обучение произношению отдельных звуков (фонем). Работа в морфологическом режиме позволяет проводить обучение слогам, звукосочетаниям или же изолированным словам.

На рис. 4.2 показана блок-схема информационной системы при работе в фонетическом режиме обучения.

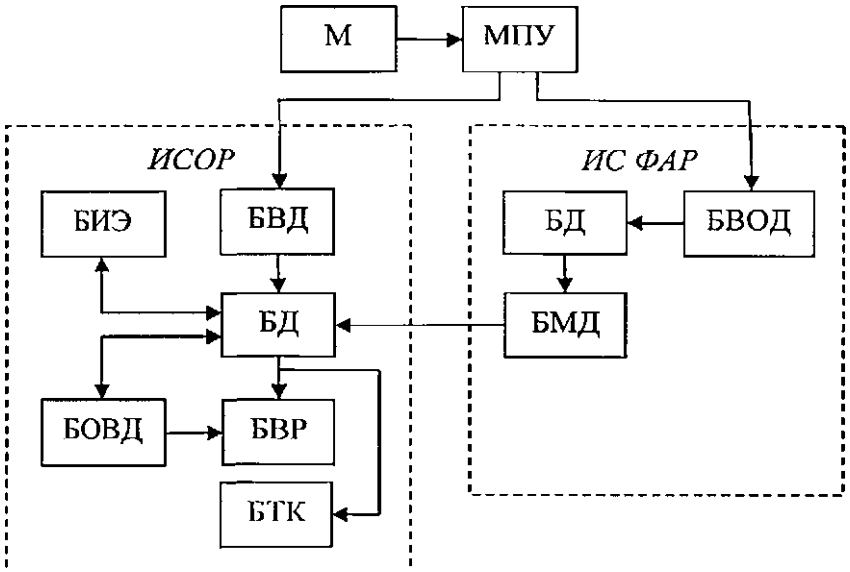


Рис. 4.2 - Архитектура ИСОР

Информационная система (ИС) для обучения языку на фонетическом уровне функционирует в следующих режимах: режиме подготовки данных (режимы А и Б), обучения и тестирования качества речи; и структурно состоит из двух частей: собственно информационной системы обучения речи (ИСОР) и ИС ФАР.

Режим подготовки данных А предназначен для записи в базу данных (БД) ИСОР эталонных реализаций фонем. В данном режиме сигнал (основные фонемы) от диктора с заведомо хорошим качеством речи и произношением поступает на динамический микрофон (М) и далее на микрофонный ламповый предусилитель со встроенным аналого-цифровым преобразователем (МПУ), который осуществляет усиление сигнала и его преобразование в цифровой вид. Оцифрованный речевой сигнал подается на вход блока ввода данных (БВД) ИСОР, в котором в автоматическом режиме происходит вначале его запись в звуковой *.WAV файл, а затем его считывание, обработка и запись в БД. Введенные в данном режиме сигналы используются:

- в блоке обработки и визуализации данных (БОВД) - в качестве ИЦ-эталонов структурированной области допустимых решений;

• в блоке идентификации эталонов (БИЭ) - в качестве обучающего словаря при распознавании с мигрированных из ИС ФАР множества ГГФБД дикторов (подробнее см. режим подготовки данных Б);

• в блоке вывода результатов (БВР) - при озвучивании эталона произношения фонем в процессе обучения.

Режим подготовки данных Б служит для заполнения БД ИСОР эталонными реализациями фонем, обработанных в ИС ФАР (на рис. 1 представлена только та часть структурной схемы ИС ФАР, которая необходима для описания режима подготовки данных Б; подробнее о структурной схеме ИС ФАР и назначении ее блоков см. [Описание ИС ФАР]). В отличие от режима А, звуковой сигнал здесь представляет собой не короткие реализации основных фонем, а отрывок устной слитной речи продолжительностью от нескольких минут. Анализируемый сигнал записывается в соответствующий *.WAV файл в память персонального компьютера с помощью микрофона М и предусилителя МПУ и далее считывается в ручном режиме блоком ввода и обработки данных (БВОД) ИС ФАР. В БВОД происходит автоматическое сегментирование речевого сигнала на короткие однородные отрезки данных — ЭРЕ и последующее их группирование в речевые кластеры с вычислением ИЦЭ. Таким образом, на выходе БВОД формируется ПФБД диктора, который и записывается в базу данных (БД) ИС ФАР. В ИС ФАР добавлен блок миграции данных (БМД) в ИСОР. Данный блок осуществляет считывание данных из БД ИС ФАР, преобразование структуры данных и последующую их миграцию в БД

ИСОР. При этом осуществляется проверка на наличие мигрируемых данных в БД ИСОР для предотвращения нарушения целостности и консистентности данных. Запуск БМД осуществляется пользователем для выбранного диктора с помощью графического интерфейса ИС ФАР (подробнее см. главу 3). Далее с мигрированные в БД ИСОР эталоны проходят этап именования или идентификации среди ИЦЭ, записанных в режиме подготовки данных А, в блоке идентификации эталонов (БИЭ). В данном блоке решается задача Я — альтернативного (по числу введенных ИЦЭ) распознавания образов

речи, и в результате с мигрированные реализации ПФБД ставятся в соответствие с эталонными реализациями фонем. Запуск процесса идентификации осуществляется пользователем с помощью графического интерфейса ИСОР после миграции данных. Введенные в данном режиме сигналы используются:

- в БОВД - в качестве эталонов фонем в режиме обучения;
- в БВР - при озвучивании эталонов, отображения их в виде отчета и экспорта во внешний файл;
- в БТК - при тестировании качества устной речи.

На рис. 4.3 показана блок-схема ИС при работе в морфологическом режиме.

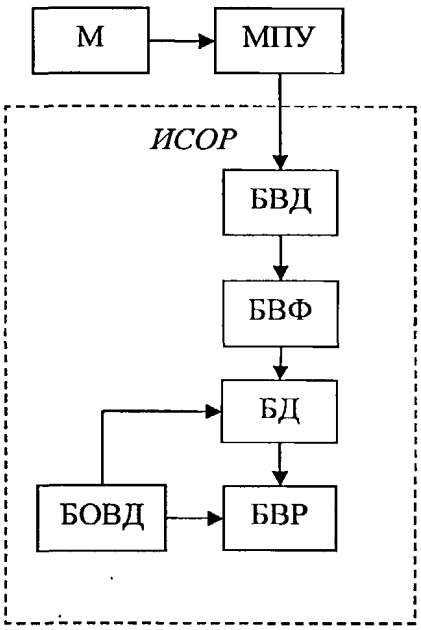


Рис. 4.3 - Архитектура ИСОР при работе на морфологическом уровне

Здесь БВФ - блок выделения фонем. В нем осуществляется выделение во входном сигнале квазистационарных участков, представление их в виде вектора АР-коэффициентов и запись полученных результатов в БД. Входной сигнал из МПУ поступает в БВД, который записывает его в звуковой файл, а затем передает полученный сигнал в БВФ. БВФ выделяет во входном сигнале фонемы и записывает их в БД, формируя, таким образом, персональную морфологическую базу данных диктора. Дальнейшее функционирование системы при визуализации результатов аналогично работе в фонетическом режиме

обучения. Несмотря на то, что названия блоков совпадают с названиями при работе в фонетическом режиме, их внутренняя структура существенно отличается.

Одним из основных режимов работы ИС является режим обучения речи и постановки произношения, заключающийся во взаимодействии через БД ИСОР блока ввода данных и блока обработки и визуализации данных. При работе в фонетическом режиме на первом этапе пользователем с помощью графического интерфейса ИСОР в БОВД осуществляется выбор одной из фонем для обучения, а так же критерии отбора эталонов. В соответствии с выбранными критериями, БОВД осуществляет запрос к БД, выбирает данные, обрабатывает и выводит их на экран ПК структурированную область допустимых решений — эталоны выбранной фонемы. Сигнал обучаемого от М через МПУ и БВД в автоматическом режиме кодируется и записывается в БД ИСОР. БОВД автоматически считывает сигнал обучаемого из БД, обрабатывает его и выводит на экран ПК. Указанная последовательность действий повторяется произвольное количество раз до достижения требуемого результата относительно качества произношения фонем.

И заключительным режимом работы ИС служит **режим тестирования качества речи.** Данный режим реализуется в блоке тестирования качества речи (БТК), который позволяет считывать звуковой файл с записью речи анализируемого диктора и сравнивать выделенные из нее фонемы с фонемами эталонного диктора из БД. Результатом работы блока является оценка качества произношения (величина, принимающая значения в диапазоне 0...1).

На рис. 4.4 показана структура ключевого звена ИСОР: блока тестирования качества речи (БТК).

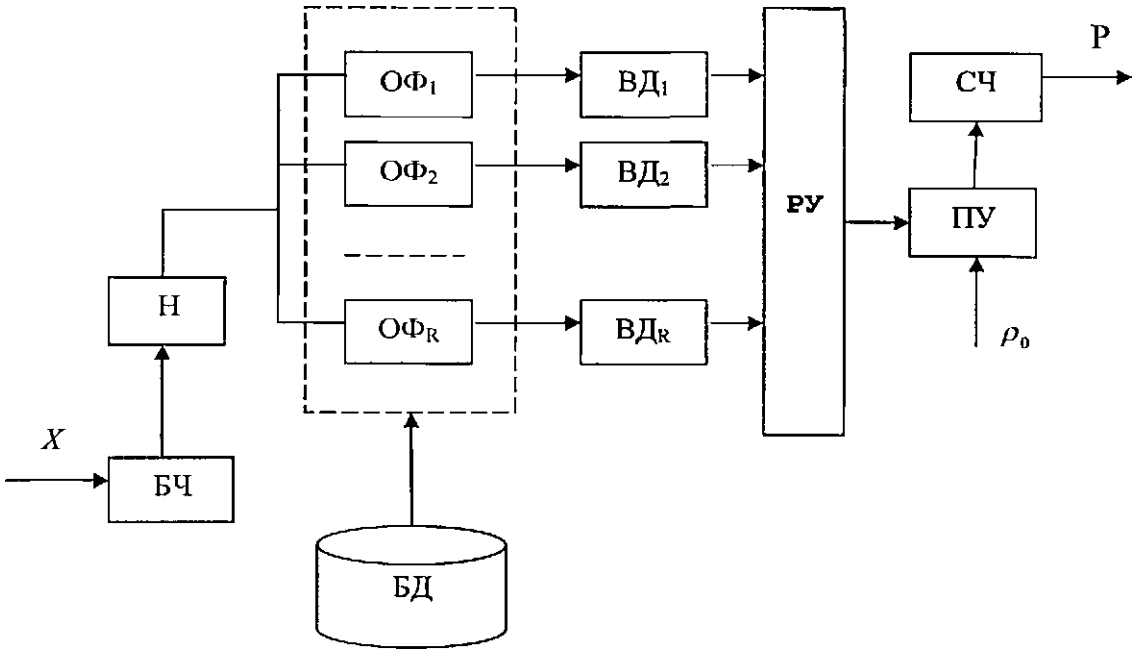


Рис. 4.4 - Внутренняя архитектура блока тестирования качества речи

Анализируемый сигнал поступает в блок чтения (БЧ), разбивающий его на сегменты длительностью 10 мс и передает их в блок нормировки Н, который нормирует каждый сегмент по дисперсии порождающего шума его АР-модели. Как показано в [76], нормировка РЕ по дисперсии порождающего шума приводит к уменьшению вариаций информационного рассогласования в пределах множества различных реализаций каждой фонемы. Полученный результат пропускается через набор обеляющих фильтров ($ОФ_1$-.$ОФ_R$). Коэффициенты ОФ $\boldsymbol{a_r, r} = \overline{1,R}$ устанавливаются в соответствии со значениями АР-коэффициентов фонем-эталонов из БД. Сигнал с выхода ОФ поступает на вычислитель дисперсии ($ВД_1$...$ВД_R$). Полученный вектор $p_r^2(X), r = \overline{1,R}$ поступает в решающее устройств РУ, которое вычисляет решающую статистику (1.10).

Разработанная информационная система использует азу данных формата Microsoft Access 2002. На рис. 4.5 показана схема этой базы данных.

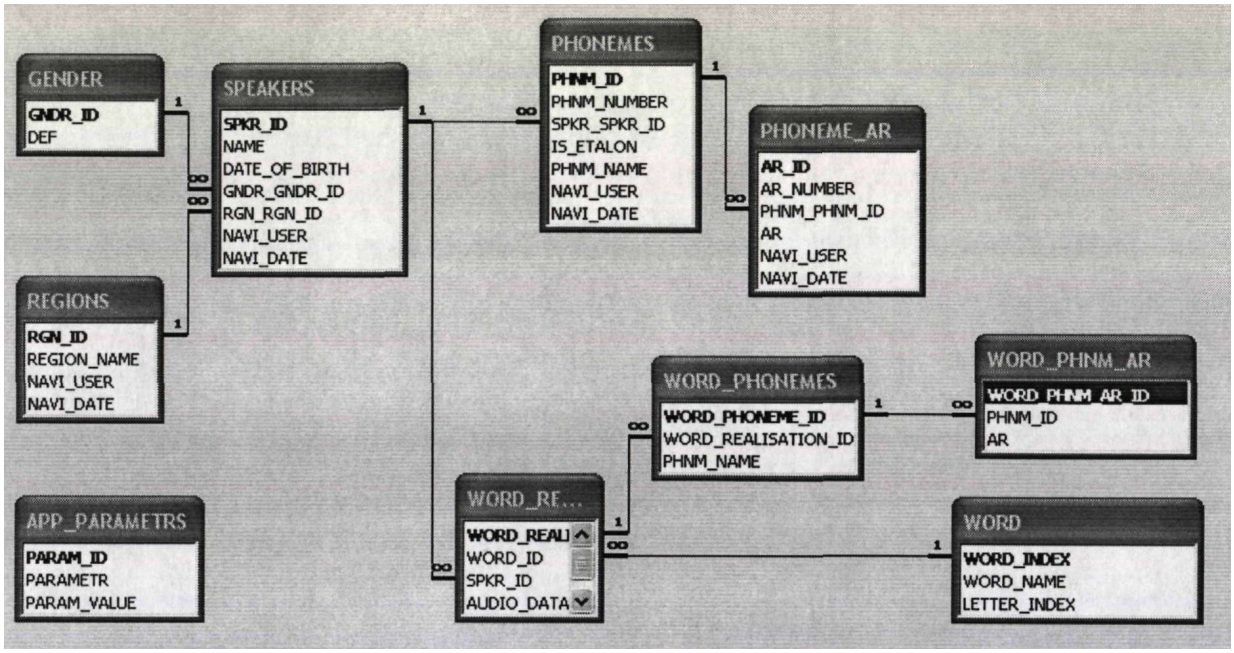


Рис. 4.5 - БД данных ИСОР

Схема включает в себя следующие таблицы:

1. GENDER - таблица-справочник, содержит определяющие пол диктора. Название полей таблицы:

- GNDR ID - идентификатор пола (первичный ключ);
- DEF - название.

2. SPEAKERS - таблица хранит личностную информацию о дикторе

- SPKR ID - идентификатор диктора (первичный ключ);
- NAME - имя диктора;
- DATE OF BIRTH - год рождения;
- GNDRGNDRID - вторичный ключ на идентификатор пола;
- RGN RGN ID - вторичный ключ на идентификатор региона;
- NAVI USER - имя пользователя;
- NAVI DATE - дата вставки DEF - название.

3.REGIONS - таблица-справочник хранит информацию о названиях регионов:

- RGN_ID - идентификатор региона (первичный ключ);
- REGION_NAME - название региона;
- NAVI_USER - имя пользователя;
- NAVI_DATE - дата вставки.

4. PHONEMES - таблица содержит информацию о фонемах:

- PHNM_ID - идентификатор фонемы (первичный ключ);
- PHNM_NUMBER - порядковый номер фонемы у диктора;
- SPKR_SPKR_ID - вторичный ключ на идентификатор диктора;
- IS_ETALON - признак эталона;
- PHNM_NAME - название фонемы;
- NAVI_USER - имя пользователя;
- NAVI_DATE - дата вставки.

5. PHONEMEAR - таблица содержит авторегрессионные коэффициенты фонем. Название полей таблицы:

- AR_ID - идентификатор AP-коэффициента (первичный ключ);
- AR_NUMBER - порядковый номер AP-коэффициента у фонемы;
- PHNM_PHNM_ID - вторичный ключ на идентификатор фонемы;
- AR - значение AP-коэффициента;
- NAVI_USER - имя пользователя;
- NAVI DATE - дата вставки.

6.APP_PARAMETRS – таблица хранит параметры и настройки ИС:

- PARAM_ID – идентификатор параметра (первичный ключ);
- PARAMETR – название параметра;
- PARAM_VALUE – значение параметра;

7. WORD - таблица хранит список слов для обучения:

- WORD_INDEX - уникальный идентификатор слова (первичный ключ);
- WORD_NAME - буквенное обозначение слова;
- LETTER_INDEX — индексное поле для использования в словаре;

8. WORD_REALISATION - таблица хранит реализации эталонов слова:

- WORD_REALISATION_ID - уникальный идентификатор реализации (первичный ключ);
- WORD_ID — идентификатор слова;
- SPKR_ID - идентификатор диктора;

- AUDIO_DATA – имя звукового файла реализации;

9. WORD_PHONEMES - таблица хранит фонемы реализации слова:

- WORD_PHONEM_ID - уникальный идентификатор фонемы (первичный ключ);
- WORD_REALISATION_ID - идентификатор реализации;
- PHNM_NAME – текстовое обозначение фонемы

10. WORD_PHNM_AR - таблица хранит список АР-коэффициентов фонемы слова:

- WORD_PHNM_AR_ID - уникальный идентификатор АР-коэффициента (первичный ключ);
- PHNM_ID - идентификатор фонемы;
- AR - значение коэффициента.

Разработанная система предназначена для обучения языку глухих и слабослышащих, постановке произношения, обучению иностранным языкам и диалектам. Программа реализована в среде разработки MATLAB 6.5. Программа позволяет:

- Проводить обучение речи и постановку произношения в двух режимах: фонетическом и морфологическом.
- Считывать речевые сигналы с микрофона, рассчитывать по ним коэффициенты авторегрессионной модели и записывать полученные результаты в БД.
- Отображать содержимое БД.
- Редактировать данные, хранящиеся в БД.
- Изменять и сохранять в БД настройки приложения.
- Экспортировать АР-модели реализаций фонем и эталонов из БД в файл.
- Формировать персональные ФБД дикторов из слитной речи
- Проводить миграцию данных из базы данных ИС ФАР в базу данных ИСОР.
- Озвучивать хранимые в БД данные.
- Осуществлять проверку качества устной речи

Программа имеет удобный и понятный графический интерфейс, реализованный средствами системы MATLAB 6.5, главная форма которого

показанный на рис. 4.6. Исходными данными для ИС служат речевые сигналы, записанные со стандартного микрофона в режиме «on-line» (в режиме подготовки данных А и режиме обучения), а также стандартные звуковые WAV-файлы (в режиме подготовки данных Б).

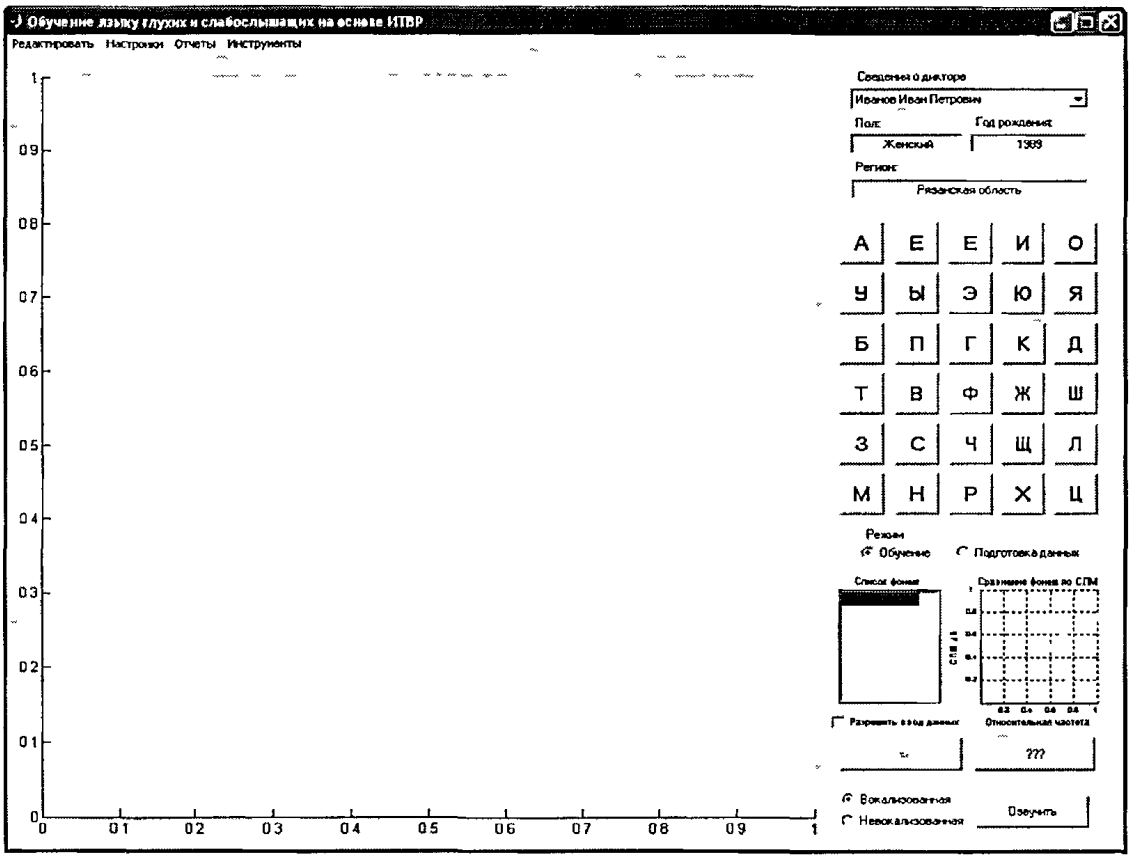


Рис. 4.6 - Вид пользовательского интерфейса ИСОР при работе на фонетическом уровне

В правом верхнем углу формы **«Обучение речи и постановка произношения на основе ИТВР»** расположена личностная информация о дикторах, ограниченная рамкой *«Сведения о дикторах»*. Выпадающий список содержит фамилии, имена и отчества дикторов; значения полей *«Пол», «Год рождения»* и «*Регион*» изменяются автоматически при выборе диктора из выпадающего списка. Ниже личностной информации дикторов расположена клавиатура, кнопки «*А*», *«Е»*, ... «Ц» которой соответствуют обучению произношения определенному звуку. При нажатии на одну из кнопок она подсвечивает красным цветом и на основном поле отображается упорядоченный

набор сигналов-эталонов, по которым и производится обучение. Ниже клавиатуры располагаются кнопки выбора *режима работы* приложения: здесь предусмотрен выбор одного из двух режимов работы ИСОР - либо *режим обучения*, либо *режим подготовки данных* (А и Б).

Ниже располагается *список*, в котором отображается набор отображенных на экране эталонов (в режиме подготовки данных) или набор реализаций фонем обучаемого (в режиме обучения). Рядом со списком отображается *график спектральной плотности мощности* (СПМ) выбранного значения списка, причем количество графиков и их цвет зависит от режима работы ИС:

- малиновый цвет - график СПМ для ИЦЭ (во всех режимах работы ИСОР);
- красный — график СПМ выделенного в списке сигнала (только в режиме обучения);
- синий цвет — график СПМ выделенного в списке эталона (режим подготовки данных) или ближайшего к сигналу от обучаемого эталона (в режиме обучения).

Расположенный под списком фонем *флаг «Разрешить ввод данных»* делает активной кнопку *«Запись»*, которая используется для ввода эталонов в режиме подготовки данных А и записи реализаций фонем в процессе обучения. Указанный флаг введен в ИС для предотвращения засорения БД случайными сигналами. Под кнопкой «Запись» находится блок озвучивания сигналов, объединенный рамкой. Воспроизведение реализаций фонем и эталонов из списка осуществляется при помощи кнопки *«Озвучить»*, при этом выбор их типа осуществляется с помощью кнопок *«Вокализованная»* и *«Невокализованная»*. На основном поле формы отображается область допустимых решений процесса обучения. На этом поле:

- синим цветом отображаются эталоны и окружности, на которых они расположены;
- малиновым цветом - ИЦЭ;
- красным — сигналы от обучаемого и траектория обучения.

В правом верхнем углы формы располагается главное меню ИСОР. Приведем краткое описание каждого из его разделов:

• Меню *"Редактировать"* содержит опции редактирования информации о дикторах и фонемах:

- *"Диктор":*
  - *«Добавить»* - позволяет добавлять нового диктора в БД (форма показана на рис. 4.7);
  - *«Изменить»* - позволяет редактировать личностную информацию диктора (форма показана на рис. 4.8);
  - *«Удалить»* - позволяет удалять диктора из БД.
- *"Регион"-.*
  - *«Добавить»* - позволяет добавлять новый регион в БД (форма показана на рис. 4.9);
  - *«Изменить»* - позволяет редактировать название региона (форма показана на рис. 4.9);
  - *«Удалить»* - позволяет удалять регион из БД.
- *"Фонемы":*
  - *«Сохранить как ИЦЭ»* - позволяет сохранить эталон выбранной фонемы в качестве информационного центра-эталона при работе в режиме подготовке данных;
  - *«Освободить ИЦЭ»* - модифицирует ИЦЭ в эталон, не удаляя его из БД (доступно в режиме подготовки данных);
  - *«Сохранить как эталон»* - позволяет сохранить сигнал от обучаемого в качестве эталонного сигнала при работе в режиме обучения;
  - *«Освободить эталон»* - модифицирует эталон в обычный сигнал, не удаляя его из БД (доступно только в режиме подготовки данных);
  - *«Удалить»* - позволяет удалять эталоны (в режиме подготовки данных) и сигналы от обучаемого (в режиме обучения) из БД.

♦ *"Идентификация эталонов"* — позволяет проводить автоматическое именование эталонов, введенных в БД в режиме подготовки данных Б.

♦ *"Разыменование эталонов"* - позволяет проводить обезличивание всех хранящихся в БД эталонов (доступно только в режиме подготовки данных).

- Меню *"Настройки"* позволяет изменять настройки приложения (см. описание формы «Настройки приложения»),
- Меню *"Отчеты"* содержит функции для вывода информации о реализациях фонем и эталонов, а так ясе личностной информации обучаемых и дикторов, их произносивших:

♦ *"Отчет по эталонам"* — открывает форму с отображением личностной информаций дикторов и АР-коэффициентами для всех эталонов текущей фонемы (см. описание формы «Отчеты»);

♦ *"Фонемы обучаемого"* — открывает форму с отображением личностной информации и АР-коэффициентами для текущей фонемы и обучаемого (см. описание формы «Отчеты»);

- Меню *"Инструменты"* содержит функции для обработки и анализа фонетической информации:

♦ *"Тестирование произношения"* — открывает форму анализа качества произношения;

♦ *"Обучение словам"* — переводит систему в режим морфологического обучения и открывает соответствующую форму. Форма «Информация о дикторе» служит для изменения личностной информации диктора/обучаемого или добавления нового. Внешний вид формы показан на рис. 4.7.

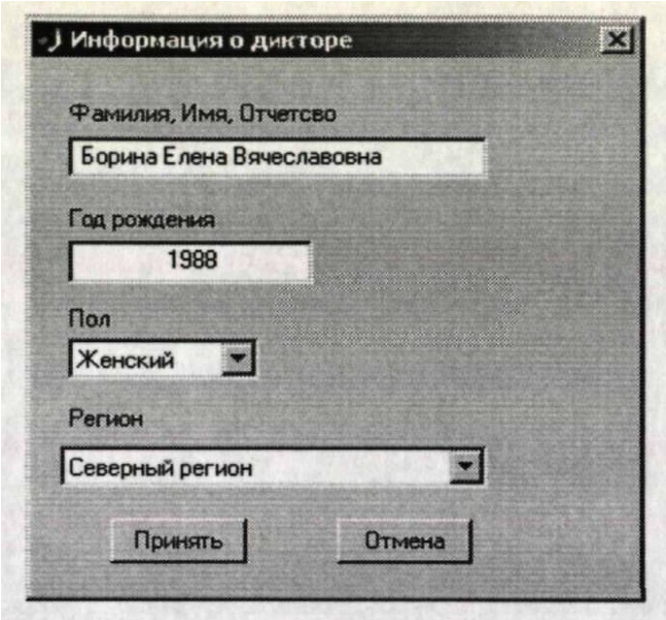


Рис. 4.7 - Форма ввода информации о дикторе

Текстовые поля формы *«Фамилия, Имя, Отчество»* и *«Год рождения»*, а так же выпадающие списки *«Пол»* и *«Регион»* служат для ввода/модификации соответствующей информации о дикторе. При этом значение списков заполняются хранимой в БД информацией и в случае необходимости модификации значения поля «Регион» необходимо воспользоваться соответствующим пунктом меню главной формы. С помощью кнопок *«Принять»* и *«Отмена»* пользователем принимается решение о сохранении/не сохранении введенной в поля формы информации.

Форма «Информация о регионе» позволяет добавлять и изменять названия регионов, хранимых в БД. Внешний вид формы показан на рис. 4.8. Текстовое поле «Название региона» служит для ввода нового региона или модификации существующего. С помощью кнопок *«Изменить»* и *«Отмена»* пользователем принимается решение о сохранении/не сохранении введенной в поля формы информации.

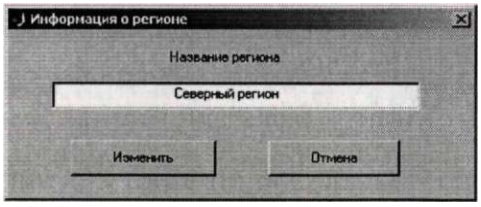


Рис. 4.8 - Форма ввода названия региона

Форма «Настройки приложения» предназначена для изменения общих настроек работы ИСОР. На рис. 4.9 представлен внешний вид данной формы. При открытии данной формы по умолчанию на ней выводятся текущие значения настроечных параметров ИСОР.

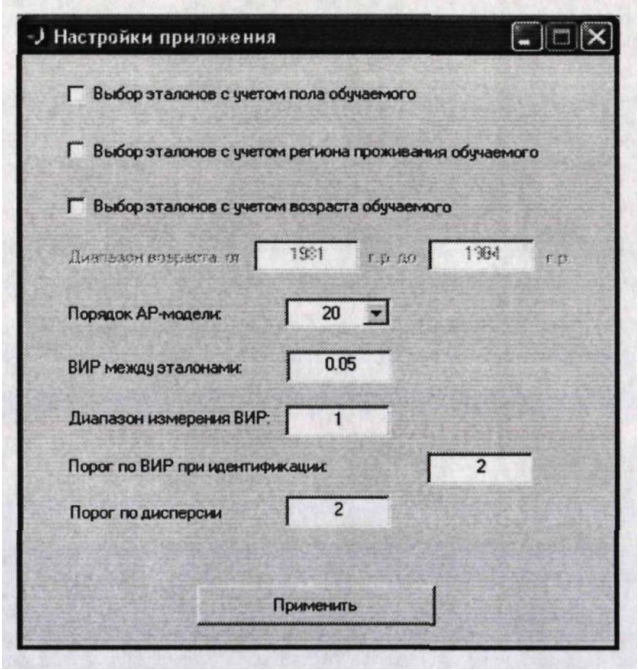


Рис. 4.9 - Форма настроек

Флаги *«Выбор эталонов с учетом пола обучаемого»*, *«Выбор эталонов с учетом региона проживания обучаемого»* и *«Выбор эталонов с учетом возраста обучаемого»* предназначены для формирования критериев отбора данных из БД. В случае необходимости учета одного из указанных критериев необходимо выставить соответствующий флаг; при этом существует возможность выставления сразу нескольких флангов. В случае выбора последнего критерия становятся активными текстовые поля для указания диапазонов изменения возрастов *«Диапазон возраста: от ... г.р. до ... г.р.»*. Выпадающий список

«Порядок АР-модели» предназначен для выбора значения данного параметра, при этом в списке содержатся значения от 5 до 30. Тестовые поля *«ВИР между эталонами»*, *«Диапазон измерения ВИР»* и *«Порог по ВИР при идентификации»* предусмотрены для изменения соответствующих параметров в настройках ИСОР. Поле *«Порог подисперсии*» предназначено для соответствующего параметра в алгоритме тестирования качества произношения. Сохранение значений введенных параметров осуществляется при нажатии на кнопку *«Применить»*.

Форма «Отчеты» позволяет выводить на экран сводную информацию по эталонам и реализациям фонем, хранимым в БД. На рис. 4.10 представлен внешний вид формы в случае выбора пункта меню «Отчет по эталонам» главной формы ИС. При открытии формы прокручивающийся *список эталонов* заполняется всеми значениями эталонов для текущего звука, хранимых в БД. При выборе одно из них на расположенном рядом графике отображаются соответствующие ему коэффициенты линейного предсказания, а в верхней части формы - личностная информация о дикторе. С помощью кнопки *«Озвучивание»* и кнопок выбора типа фонем *«Вокализованная»* и *«Невокализованная»* осуществляется воспроизведение выбранного в списке эталона. В случае если эталон введен в режиме подготовки данных А - воспроизводится соответствующий звуковой файл (кнопки выбора типа фонем не влияют на качество воспроизведения), а если эталон с мигрирован из ИС ФАР, то воспроизводится синтезированный по АР-коэффициентам звуковой сигнал. Командная кнопка *«Экспорт»* предназначена для экспорта всех отображенных в списке эталонов во внешний файл.

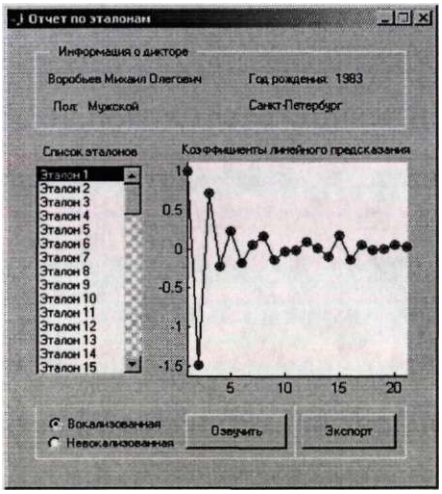


Рис. 4.10 - Форма отчета по эталонам

Форма **«Тестирование произношения»** позволяет проверить качество произношения обучаемого путем сравнения с произношением некоторого эталонного диктора. Данная форма показана на рис. 4.11.

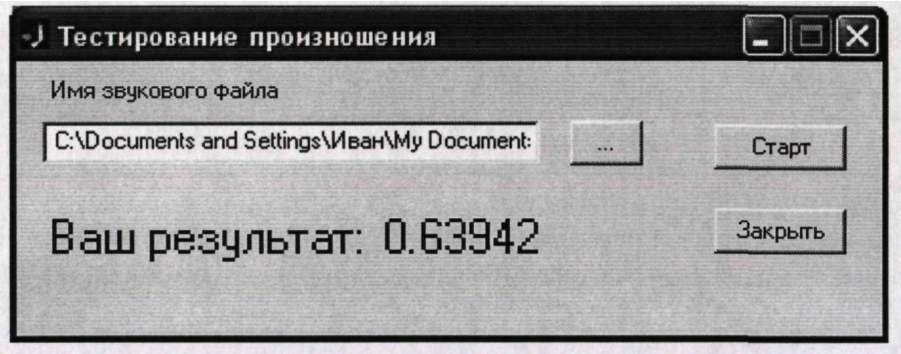


Рис. 4.11- форма тестирования качества речи

Для того чтобы оценить качество своего произношения, диктор должен записать в звуковой файл прочитанный им небольшой отрывок текста длительностью 1 - 2 минуты. Имя полученного файла необходимо ввести в соответствующее поле. После этого с помощью кнопки «Старт» запускается процедура оценки качества произношения. Ее результатом является число в диапазоне 0...1, которое отображается в поле «Ваш результат».

Форма **«Обучение словам»** позволяет проводить обучение на морфологическом уровне: изолированным словам и слогам, а также их сочетаниям. Ее общий вид показан на рис. 4.12.

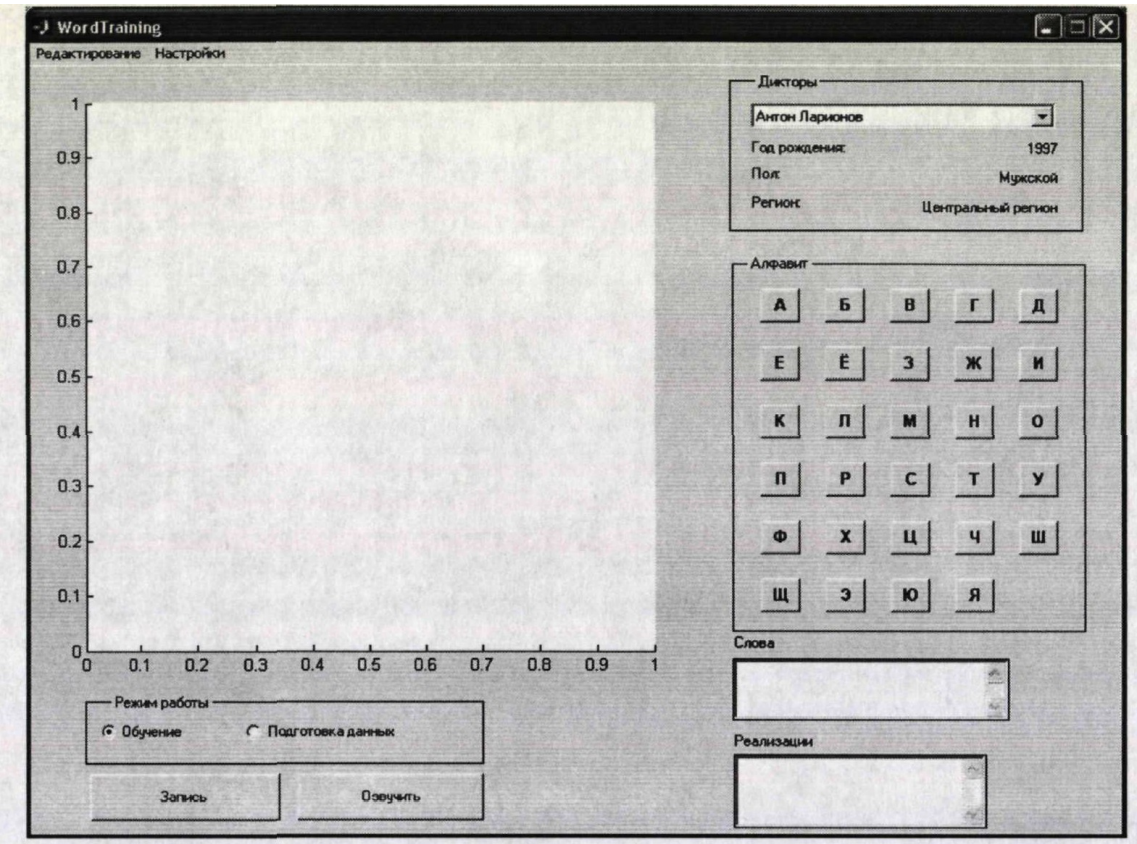


Рис. 4.12 - Вид пользовательского интерфейса ИСОР при работе на морфологическом уровне

Структура данной формы во многом аналогична структуре формы обучения фонемам, показанной на рис. 4.6. В правом верхнем углу располагается личностная информация о дикторах, ограниченная рамкой *«Дикторы»*. Выпадающий список содержит фамилии, имена и отчества дикторов; значения полей *«Пол»*, *«Год рождения»* и «*Регион*» изменяются автоматически при выборе диктора из выпадающего списка. Ниже личностной информации дикторов расположена клавиатура, кнопки *«А»*, *«Б»*, ... «Я» которой позволяют выбрать букву, с которой будут начинаться слова для обучения. При нажатии на одну из кнопок она подсвечивается красным цветом и в списке «Слова» отображаются

доступные для обучения слова. При выборе какого-либо слова на основном поле отображается упорядоченный набор сигналов-эталонов, по которым и производится обучение. Под основным полем располагаются кнопки выбора *режима работы* приложения: здесь предусмотрен выбор одного из двух режимов работы ИСОР — либо *режим обучения*, либо *режим подготовки данных.* Под списком слов располагается список «Реализации», который содержит набор отображенных на экране эталонов (в режиме подготовки данных) или набор реализаций слов обучаемого (в режиме обучения). Под рамкой выбора режима работы располагаются кнопки *«Запись»* и *«Озвучить»*. Кнопка *«Запись»* используется для ввода эталонов в режиме подготовки данных и записи реализаций слов в процессе обучения. Воспроизведение записанных реализаций слов и эталонов из списка осуществляется при помощи кнопки *«Озвучить»*. На основном поле формы отображается структурированная область допустимых решений процесса обучения, при этом:

- синим цветом отображаются на поле эталоны и окружности, на которых они расположены;
- малиновым цветом - ИЦЭ;
- красным — сигналы от обучаемого и траектория обучения.

В правом верхнем углы формы располагается главное меню ИСОР. Краткое описание каждого из его разделов:

- Меню *"Редактирование"* содержит опции редактирования информации о словах и их реализациях:
  - *"Слово":*
    - *«Добавить*» - позволяет добавить новое слово, начинающееся на выбранную букву;
    - *«Удалить»* - удаляет выбранное слово из списка;
    - *«Удалить все*» - удаляет из списка все слова, начинающиеся на заданную букву;
    - *«Список*» - выводит список всех введенных слов;
  - *"Эталон"*

■ *«Сделать ИЦу>* - позволяет сохранить эталон выбранного слова в качестве ИЦ-эталона при работе в режиме подготовки данных;

■ *«Освободить ИЦ»* - модифицирует ИЦЭ в эталон, не удаляя его из БД (доступно в режиме подготовки данных);

■ *«Удалить»* - позволяет удалять эталоны (в режиме подготовки данных) и сигналы от обучаемого (в режиме обучения) из БД;

■ *«Сегментирование»* - выводит на экран результаты сегментирования выбранного эталона в режиме подготовки данных или сигнала от обучаемого в режиме обучения;

♦ *"Выход"* - завершение работы в режиме обучения на морфологическом уровне и возврат к главной форме. • Меню *"Настройки"* позволяет изменять настройки приложения (см. описание формы «Настройки»).

Форма «Список слов» служит для добавления или изменения общего списка слов для обучения. Внешний вид формы показан на рис. 4.13.

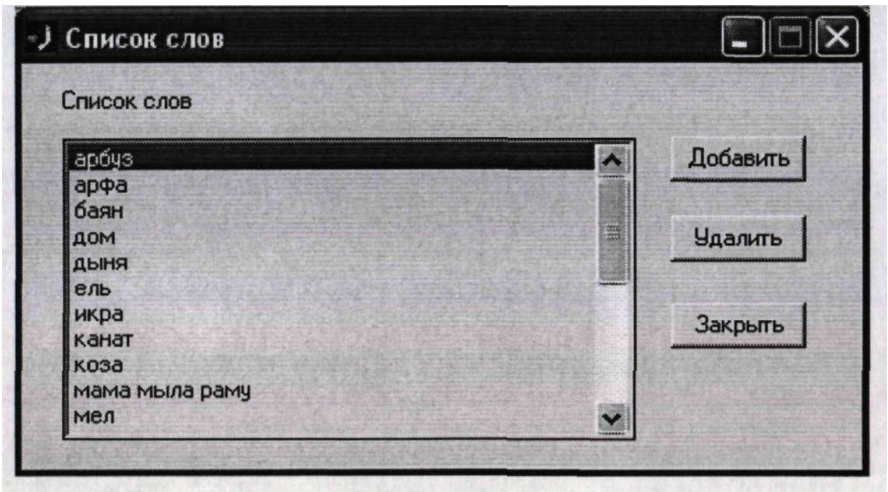


Рис. 4.13 - Форма работы со списком слов

С помощью данной формы пользователь может добавлять или удалять слова из списка, нажимая соответствующие кнопки («Добавить» и «Удалить»). При добавлении слова в список на экране появится форма «Добавить слово», внешний вид которой показан на рис. 4.14.



Рис. 4.14 - Форма добавления нового слова

Для того чтобы удалить какое-либо слово, необходимо выделить его в списке и нажать кнопку «Удалить». Стоит отметить, что работа со списком слов также возможна через контекстное меню, которое вызывается при нажатии правой кнопки на списке «Слова». Однако при этом пользователь может добавлять только те слова, которые начинаются на букву, выбранную на кнопочной клавиатуре.

Форма «Настройки» предназначена для изменения настроек работы ИСОР в режиме обучения на морфологическом уровне. На рис. 4.15 представлен внешний вид данной формы. При открытии данной формы по умолчанию на ней выводятся текущие значения настроечных параметров ИСОР.

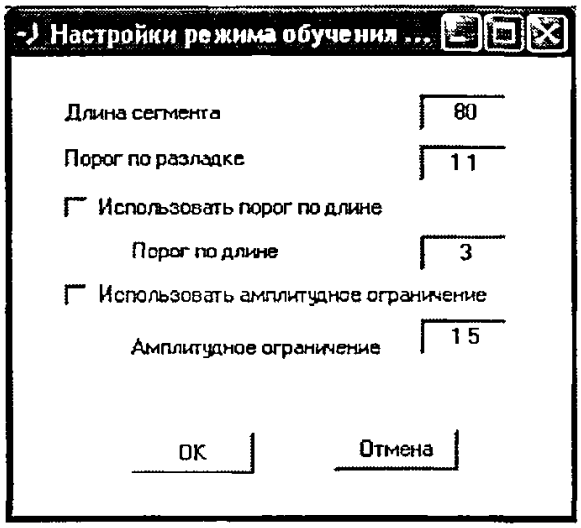
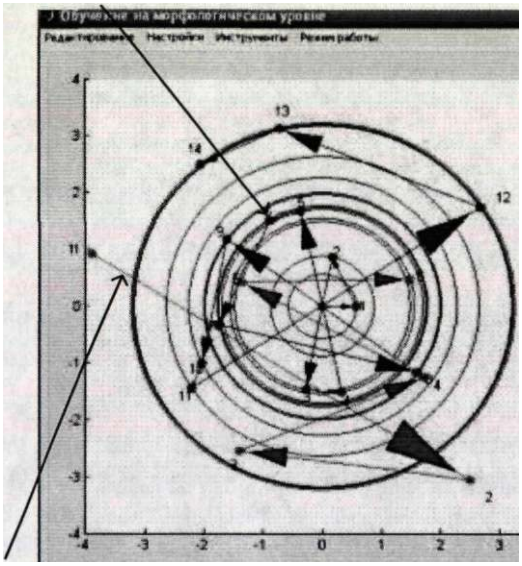


Рис. 4.15 - Форма настроек режима обучения на морфологическом уровне

Здесь поле «*Длина сегмента*» определяет длину сегмента в отсчетах, которая будет использоваться при обработке входных сигналов. Поле «*Порог по*

Эталоны



при

о

разладке» задает пороговое значение информационного рассогласования между квазистационарным участком и следующим за ним сегментом, котором принимается решение наличии разладки. Флажок *«Использовать порог по длине*» определяет, будет ли использоваться ограничение на минимальную длину фонемы. Поле «Порог по длине» определяет минимальную длину фонемы в сегментах. Если выделенная фонема оказывается меньше этого порога, то она не учитывается при дальнейшей обработке. Таким образом, если установлен флажок *«Использовать порог по длине*», то значение порога по длине считается равным нолю, т.е. учитываются все найденные в слове фонемы. Флажок *«Использовать амплитудное ограничение*» определяет, будет ли использоваться ограничение на величину информационного рассогласования при сравнении друг с другом реализации какого-либо слова. Поле *«Амплитудное ограничение*» определяет величину этого ограничения.

Выводы по главе

Благодаря применению кластерной модели РЕ была разработана новая технология информационного обеспечения в задаче обучения речи людей с дефектами в органах слуха. На рис. 4.20 представлена схема учебного процесса на ее базе.

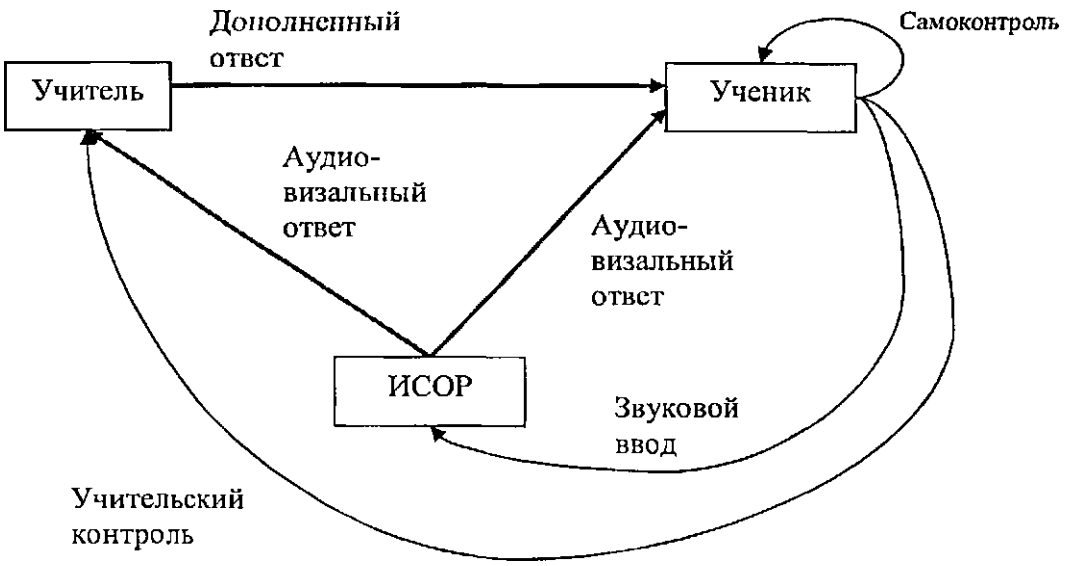


Рис. 4.20 - Схема обучения произношению при помощи разработанной технологии

Особенностью данной технологии является новый метод визуализации данных в информационной метрике Кульбака-Лейблера, основанный на двух разновидностях кластерных моделей речи: фонетической и морфологической. При смене этих двух моделей кардинально не меняется ни интерфейс, ни схема обучения. Меняется лишь содержание эталонной базы данных: вместо отдельных звуков-фонем в нее войдут отдельные звукосочетания, целые слова и даже короткие фразы. Критерий МИР (1.7) в принципиальном отношении также не изменится: в нем в качестве интегрального показателя величины информационного рассогласования по каждому слову целиком используется суммарная величина ИР (1.14) по группе всех фонем слова [106]. При этом дополнительно применяется лишь стандартная процедура выравнивания темпа речи диктора-обучаемого и диктора-эталона [83].

Было показано, что обучение глухих детей произношению по разработанной технологии позволяет им улучшить собственно произношение. Об этом наглядно свидетельствуют показатели качества речи каждого из них. У всех обучаемых уровень нечеткости речи (процент ошибок) к концу обучения снизился на 10 — 20%.

Рассмотренная технология может использоваться не только при обучении речи глухих и слабослышащих, но и при решении задачи обучения иностранным языкам и диалектам. Одна из основных проблем при изучении иностранного языка заключается во влиянии звуков национального языка на произношение звуков изучаемого языка. Все это оказывает негативное влияние, как на восприятие, так и на артикуляцию. Разработанная система может быть успешно использована при решении данной проблемы. Кроме того, с помощью режима тестирования качества речи обучаемый может контролировать произношение не только отдельных звуков, но так же и общее качество своей речи.

Заключение

Рассмотренный теоретико-информационный подход в задачах фонетического анализа речи, по-видимому, не имеет серьезных альтернатив ввиду острейшей проблемы вариативности устной речи. Данная проблема заключается, прежде всего, в особенностях речевого механизма человека. У разных людей речевой механизм сильно разнится по своим параметрам, например, по частотным характеристикам его модели «акустической трубы». Последняя, кроме того, может сильно варьироваться под действием целого ряда факторов: времени суток, эмоционального состояния диктора и т.п. Как результат, даже одноименные речевые метки от одного и того же диктора, разнесенные между собой во времени и в пространстве, могут иметь существенно разные автокорреляционные (статистические) свойства. И здесь логично возникает идея объединения близких друг другу по своему звучанию, однородных в теоретико-информационном смысле элементарных речевых единиц в соответствующие фонемы-кластеры. Это базовое положение информационной теории восприятия речи.

Благодаря проведенным исследованиям в диссертации получены следующие основные результаты:

1. Разработан новый алгоритм распознавания речи, особенностью которого является применение кластерной модели элементарных речевых единиц, критерия минимального информационного рассогласования и амплитудного ограничения на величину решающей статистики.
2. Проведены экспериментальные исследования предложенного алгоритма распознавания речи. Показано, что данный алгоритм характеризуется значительным (в 2 - 5 раз) выигрышем по своей вычислительной сложности по сравнению с известными аналогами.
3. На основе критерия минимального информационного рассогласования разработана адаптивная кластерная модель элементарных речевых единиц в

информационной метрике Кульбака-Лейблера, формирование которой происходит непосредственно во время анализа потока речи. Ее точность и надежность регулируются выбором двух основных параметров: порога по величине информационного рассогласования и порога по длительности элементарных речевых единиц. С учетом эффекта насыщения, даны экспериментальные оценки их оптимальных значений. Показано, что эти оценки распространяются на разных дикторов.

4. В лабораторных и натурных условиях проведено экспериментальное исследование адаптивной кластерной модели элементарных речевых единиц. Создана кластерная фонетическая база данных на основе речи жителей нижегородского языкового ареала. Показано, что наибольшие фонетические различия в произношении наблюдаются у жителей севера и юга Нижегородской области.
5. На основе разработанной адаптивной кластерной модели элементарных речевых единиц и критерия минимального информационного рассогласования создана новая технология информационного обеспечения при обучении речи и постановке произношения у глухих и слабослышащих лиц.
6. Для апробации и исследований предложенной технологии был разработан лабораторный макет информационной системы обучения речи с визуализацией речевого сигнала по критерию минимального информационного рассогласования. Исследования в натурных условиях подтвердили высокую эффективность разработанной технологии.

На основании полученных результатов можно сделать следующие выводы:

1. Использование ограничения в решающей статистике позволяет учитывать только информативные сегменты, что повышает надежность работы систем автоматического распознавания речи.
2. Благодаря направленным свойствам решающей статистики МИР можно значительно сократить вычислительные затраты в системах автоматического распознавания речи.

3. Разработанная адаптивная кластерная модель элементарных речевых единиц может найти широкое применение при создании новых и обновлении существующих речевых баз данных.
4. Разработанная информационная система обучения речи может найти практическое применение не только при постановке речи и обучении языкам, но и в исследовательских целях, например, при сравнительном анализе качества произношения разных дикторов, анализе зависимости их произношений от внешних факторов, в частности, эмоционального состояния и т.п.

Из сделанных выводов следует, что предложенные в диссертационной работе алгоритмы на основе кластерной модели элементарных речевых единиц могут иметь практическое применение при решении широкого круга актуальных задач в области анализа, распознавания и обработки речи. Например, это задача разработки разнообразных речевых корпусов в пределах национального языка - как от отдельных его носителей, так и их групп, объединенных по признакам пола, года рождения, места жительства и т.п. с автоматическим наполнением и непрерывным развитием (модернизацией) во времени. На этом примере проявляются все основные преимущества кластерной модели элементарных речевых единиц: высокая чувствительность к рассогласованию данных, способность к эффективной выборочной адаптации, минимум вычислительных затрат и требований к объему запоминающих устройств и другие. Сама идея информационного центра-эталона речевого образа здесь приобретает очевидный практический смысл и дополнительное обоснование.

К числу приоритетных направлений практического применения принадлежит также область современной прикладной лингвистики в части теории и практики преподавания иностранных языков. Разработанная информационная система обучения речи позволяет существенно расширить арсенал используемых здесь технических и программных средств, особенно для самостоятельной работы учащихся.

Список литературы

1. Ли У.А. и др. Методы автоматического распознавания речи: В 2-х книгах. Пер. с англ. /Под ред. У. Ли. - М.: Мир, 1983. - Кн. 1. 328 е., ил.
2. Потапова Р. К. Речь: коммуникация, информатика, кибернетика. М.: Радио и связь, 2003. 563 с.
3. Rabiner L. A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proceedings of the IEEE, 1989, Vol. 77 no. 2, pp. 257285.
4. Винцюк Т.К. Распознавание слов устной речи методом динамического программирования. // Кибернетика, 1968, № 1, с. 81 — 88.
5. Sakoe, H., Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 26, 1978, pp. 43-49.
6. M.M. Hochberg, L.T. Niles, J.T. Foote, and H.F. Silverman, "Hidden Markov Model/Neural Network Training Techniques for Connected Alphadigit Speech recognition" in Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 109-112, Toronto, Ontario, Canada, April 1991.
7. P. Haffner, M. Franzini, and A. Waibel, "Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition," in Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 105-109, Toronto, Ontario, Canada, April 1991.
8. J. Picone, "Continuous Speech Recognition Using Hidden Markov Models," IEEE ASSP Magazine, vol. 7, no. 3, pp. 26-41, July 1990.
9. Галунов В.И., Соловьев А.Н. Современные проблемы распознавания речи // Информационные технологии и вычислительные системы, 2004, №2.
10. Минский М. Л., Пейперт С. Персептроны. -М.: Мир. - 1971.
11. Головко В. Нейроинтеллект: Теория и применения. Книга 1. Организация и обучение нейронных сетей с прямыми и обратными связями - Брест: БПИ, 1999, — 260С.
12. В.И.Левенштейн, Двоичные коды с исправлением выпадений, вставок и замещений символов, Докл. АН СССР, 163, 4, 1965, 845-848. р.

13. Baum L.E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. // Ann. Math. Stat. Vol. 37, pp. 1554 - 1563, 1966.
14. Baum L.E., Egon J.A. An inequality with application to statistical estimation for probabilistic functions of Markov process and to a model for ecology. // Bull. Amer. Meteorol. Soc., vol. 73, pp. 360 - 363, 1967.
15. Baum L.E., Sell G.R. Growth functions for transformations on manifolds. // Pac. J. Math., vol. 27, no. 2, pp. 211 - 227, 1968.
16. Baker J.K. "Stochastic modeling for automatic speech understanding" in Speech Recognition, D.R. Reddy, Ed. New York: Academic Press, 1975, pp.
. 521 -542.
17. Baker J.K. The DRAGON system - An overview. // IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-23, pp. 24 - 29, 1975.
18. Lowerre B.T. The HARPY speech understanding system. Unpublished Ph.D. dissertation, Dept. of Comput. Sci., Carnegie-Mellon Univ., Pittsburg, PA, 1976.
19. Liporace L.R. Maximum likelihood estimation for multivariate observations of Markov sources. // IEEE Trans. Inform. Theory, IT-28:729 - 734, September 1982.
20. Juang B.H., Levinson S.E., Sondhi S.M. Maximum likelihood estimation for multivariate mixture observations of Markov chains. // IEEE Trans. Inform. Theory, Vol. IT-32, No.2, pp.307-309, Mar. 1986.
21. Левинсон С.Е. Структурные методы автоматического распознавания речи. // ТИИЭР, т. 73, № 11, 1985. с. 100 - 128.
22. Потапова Р.К. Основные тенденции многоязычной корпусной лингвистики. // Речевые технологии. № 2, 2009. с. 92 — 114.
23. Е.В. Бенедиктова, Д.Н. Лавров Идентификация диктора по фиксированному набору частот с помощью линейного классификатора. // Математические структуры и моделирование. 2008, вып. 18, с. 108-115.
24. Campbell J. P. Speaker recognition: a tutorial. Proc. IEEE. Vol. 85, N. 9. 1997. P. 1437-1462.
25. B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE,* vol. 64, pp. 460—475, 1976.

26. G. R. Doddington, "Speaker recognition—Identifying people by their voices," *Proc. IEEE,* vol. 73, pp. 1651-1664, Nov. 1985.
27. S. Furui, "Speaker-dependent-feature extraction, recognition and processing techniques," *Speech Commun.,* vol. 10, pp. 505—520, 1991.
28. A. E. Rosenberg and F. K. Soong, "Recent research in automatic speaker recognition," in *Advances in Speech Signal Processing,* S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 701-738.
29. Rossiter D., Howard D.M., Downes M. "A realtime LPC based vocal tract area display for voice development", 1993, Voice Foundation's 22nd Anniversary Symposium: Care of the Professional Voice, June, Philadelphia, Pennsylvania.
30. Benedetto M.D., Destombes F., Merialdo B., and Tubach J.P. "Phonetic recognition to assist lipreading for deaf children", 1982, Proc. IEEE ICASSP- 82.
31. Kewley - Port D., Watson C.S., and Cromer P.A "The Indiana Speech Training Aid ISTRA: A microcomputer-based aid using speaker-dependent speech recognition". Synergy '87, The 1987 ASHF Computer Conference, Proceedings, pp. 94 - 99.
32. Kewley-Port D., Watson C.S., Elbert M., Maki K., Reed D. "The Indiana Speech Training Aid ISTRA II : training curriculum and selected case studies", 1991, Clinical Linguistics and Phonetics, vol. 5.
33. Watson C.S., Kewley-Port D. "Advances in Computer-based speech training (CBST): Aids for the profoundly hearing impaired", in Research on the Use of Sensory Aids for Hearing-Impaired Persons, N. McGarr, (Ed.), Volta Review, 91(4), 29-45, 1989.
34. Fitzgerald M., Gruenwald A., Stoker R., "Software review - Video Voice Speech Training System", 1989, Review, vol. 89, pp. 171-173.
35. Rodellar V., Nieto V., Gomez P., Martinez D., and Perez M. (1994), "A Neural Network for Phonetically Decoding the Speech Trace". Proc. 1994 International Conference on Spoken Language Processing (ICSLP94) pp 1575 - 1578.
36. Rooney E., Carraro F., Dempsey W., Robertson K., Vaughan R., Jack M., Murray J. "HARP - An autonomous speech rehabilitation system for hearing- impaired

people", 1994, Proc. 1994 International Conference on Spoken Language Processing (ICSLP94) pp 2019 - 2022.

37. Rooney E., Jack M., Lefevre J., and Sutherland A. "HARP - A speech training aid for the hearing impaired", 2nd TIDE Congress, La Villette, Paris, 26th- 28th April 1995.
38. Аграновский А.В., Леднов Д.А., Карпов И.А., Маетная Р.И., Черкашина Е.Ю. Контроль произношения глухих и слабослышащих детей // Учебно-методическая конференция «Современные информационные технологии в учебном процессе» ч.2, тезисы докладов 26-27 апреля 2001г., РГУ, Ростов-на-Дону, стр.3-4.
39. A. Neri, C. Cucchiarini, H. Strik. Feedback in Computer Assisted Pronunciation Training: technology push or demand pull? Proceedings of ICSLP 2002, Denver, USA, pp. 1209-1212.
40. Потапов А.С. Распознавание образов и машинное восприятие: Общий подход на основе принципа минимальной длины описания. - СПб.: Политехника, 2007. - 548 с.
41. Нильсон Н. Обучающиеся машины, «Мир», 1967.
42. Meisel W. Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York, 1972.
43. Narasimhan R. A Linguistic Approach to Pattern Recognition // Rep. 121, Digital Computer Lab., Univ. of Illinois, Urbana, 1962.
44. Miller W.F., Shaw A.C. Linguistic methods in picture processing - A survey // Proc. AFIPS Fall Joint Comput. Conf. 1968. - pp. 279 - 290.
45. Завалишин Н.В., Мучник И.Б. Лингвистический (структурный) подход к проблеме распознавания образов. // Автоматика и телемеханика. - 1969, № 8. сс. 86-118.
46. Фу К. Структурные методы в распознавании образов. М.: Мир, 1977. - 320 с.
47. Tyron R.C. Cluster Analysis: Ann Arbor, MI, Edwards Brothers, 1939.
48. Фукунага К. Введение в статистическую теорию распознавания образов. Пер. с англ. -М.: Наука. 1979, 368 с.

49. Савченко В.В.. Информационная теория восприятия речи // Известия вузов. Радиоэлектроника. 2007. Вып.6. С. 10-14.
50. Кульбак С. Теория информации и статистика. М.: Наука, 1967. 408 с.
51. Савченко В.В. Фонема как элемент информационной теории восприятия речи. // Известия вузов России. Радиоэлектроника. 2008. Вып.4.
52. Савченко В.В. Теоретико-информационное обоснование гауссовой модели сигналов в задачах автоматической обработки речи. // Известия вузов России. Радиоэлектроника. 2008. Вып.1. с. 24-33.
53. Rabiner L.R., Juang B.-H. Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ, 1993.
54. Komanski R., Macukow B. Problems Connected with Application of Neural Networks in Automatic Face Recognition // ICAISC 2004, LNAI 3070, pp. 736-741, 2004.
55. Dempster A.P. Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. // J. Roy. Stat. Soc., vol. 39, no. 1, pp. 1 - 38, 1977.
56. Zica Valsan, Inge Gavat, Bogdan Sabac и др. Statistical and Hybrid Methods for Speech Recognition in Romanian // International Journal of Speech Technology 5, 259-268, 2002.
57. Винцюк, Т. К. Организация вычислений при распознавании больших словарей // Автоматическое распознавание и синтез речевых сигналов: Сб. науч. тр. Киев, 1989.
58. Nuance speech recognition system, Version 8.0. Introduction to the Nuance system. / Nuance Communications, Inc. 2001. 122 p.
59. Helmholtz H. Die Lehre von der Tonempfindungen als physiologische Graudlage fur die Theorie der Musik, Brounschweig, 1870.
60. Chiba T., Kajiama M. The vowel, its nature and structure. - Tokyo, 1941.
61. Фант Г. Акустическая теория речеобразования. — М.: Наука, 1964.
62. Kent R.D. at al.(Eds) Papers in Speech Communication: Speech Production, Ac. Soc. of America, 1991.

63. Sondhi M.M. Model for wave propagation in a lossy vocal tract. J. Acoust. Soc. Amer., 55: 1070 - 1075, 1974.
64. Portnoff M.R. A quasi-one-dimensional digital simulation for the time varying vocal tract. Master's thesis, MIT, 1973.
65. Dunn H.K. The Calculation of Vowel Resonances and an Electrical Vocal Tract. J. Acoust. Soc. Amer., 22: 740 - 753, 1950.
66. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов: Пер. с англ. / Под ред. М.В. Назарова и Ю.Н. Прохорова. - М.: Радио и связь, 1981.
67. Makhoul J. Linear prediction: A Tutorial Review. Proc. IEEE, vol. 63, pp. 561 -580, April 1975.
68. Makhoul J. Lattice Methods in Spectral Estimation, in Applied Time Series II, D.F. Findley, ed., Academic Press, Inc., New York, 1981. pp. 301 - 324.
69. Маркел Дж. Д., Грей А.Х. Линейное предсказание речи: Пер. с англ. / Под ред. Ю.Н. Прохорова и В.С. Звездина. - М.: Связь, 1980.
70. Савченко В.В., Акатьев Д.Ю. Авто тестирование качества произношения речи по принципу минимального информационного рассогласования // Сборник научных трудов «Современные тенденции компьютеризации процесса изучения иностранных языков». Луганск: Восточно-украинский национальный университет. 2005. Вып.3. С.205-206.
71. Савченко В.В. Автоматическая обработка речи по критерию минимума информационного рассогласования на основе метода обеляющего фильтра // Радиотехника и электроника. 2005. Т50. №3. С.309-314.
72. Савченко В.В., Акатьев Д.Ю., Шерстнев С.Н. Метод оптимального обучающего словаря в задаче распознавания речевых сигналов по критерию минимального информационного рассогласования // Известия вузов. Радиоэлектроника. 2006. Вып.5. С. 10-14 .
73. Picone J. Signal Modeling Techniques in Speech Recognition Proceedings of the IEEE, vol.81, no 9, pp. 1215-1246, Sept. 1993.

74. Савченко В.В. Автоматическое распознавание речи методом .дерева на основе информационного (i? +1)-элемента // Изв. вузов России. Радиоэлектроника. 2006. Вып.4. С. 13 -22.
75. Савченко В.В. Различение случайных сигналов в частотной области // Радиотехника и электроника. 1997. Т.42. №4. С. 426-431.
76. Савченко В.В., Акатьев Д.Ю., Карпов Н.В. Автоматическое распознавание речевых единиц методом обеляющего фильтра. // Известия вузов. Радиоэлектроника. 2007. Вып.4. С.11-19.
77. Савченко В.В. Метод фонетического декодирования слов в задаче автоматического распознавания речи на основе принципа минимума информационного рассогласования. // Известия вузов России. Радиоэлектроника. 2009. Вып.5. с. 41 -49.
78. Марпл С.Л.-мл. Цифровой спектральный анализ и его приложения: Пер. с англ. М.: Мир, 1990. 584с.
79. Сорокин В.Н. Теория речеобразования. - М.: Радио и связь, 1985.
80. Фланаган Дж.Л. Анализ, синтез и восприятие речи. - Связь., М., 1968.
81. Levinson, S.C. Mathematical models for speech technology. Chichester, England: John Wiley & Sons Ltd, 2005, 261p.
82. Потапова Р.К. Речь: коммуникация, информация, кибернетика: Учебное пособие: Изд. 2-е, доп. М.: Эдиториал УРСС, 2001.
83. Савченко В.В. Акатьев Д.Ю. Губочкин И.В. Автоматическое распознавание изолированных слов методом обеляющего фильтра // Изв. высших учебных заведений России. Радиоэлектроника. 2007. Вып. 5. С. 11-18.
84. Финк Л. М. Теория передачи дискретных сообщений. М.: Советское радио, 1970, 727с.
85. Levy C., Linares G., Nocera P., Bonastre J.-F. Reducing computational and memory cost for cellular phone embedded speech recognition system. IEEE International conf. on Acoust., Speech and Signal Processing. 2004. vol. 5, pp. 309-312.

86. Винцюк Т.К. Сравнение ИКДП- и НММ-методов распознавания речи. // Методы и средства информатики речи. Киев, 1991. с. 4 - 9.
87. Савченко А. В. Метод направленного перебора словаря в задаче автоматического распознавания речи на основе информационной теории восприятия // Сист. упр. и инф. технол. 2009. № 1. С. 44-54.
88. Drake A.W. Discrete-state Markov processes. - Chapter 5 in Fundamentals of Applied Probability Theory. New York, NY: McGraw-Hill, 1967.
89. Viterbi A.J. Error bounds for convolutional codes and asymptotically optimal decoding algorithm. // IEEE Trans. Informat. Theory, vol. IT-13, pp. 260 — 269, Apr. 1967.
90. Савченко В.В., Акатьев Д.Ю. Автоматическое распознавание случайных сигналов по критерию минимального информационного рассогласования с переспросом. // Известия вузов России. Радиоэлектроника. 2006. Вып.1. С. 20-29.
91. Губочкин И.В. Автоматическое распознавание изолированных слов с верификацией диктора на основе критерия минимума информационного рассогласования. Материалы конференции Информационные системы и технологии ИСТ - 2009. Н. Новгород, НГТУ им. Р.А. Алексеева, 2009. с. 319-320.
92. Акатьев Д.Ю., Губочкин И.В. Распознавание речи по критерию минимума информационного рассогласования с обнаружением ошибок. Материалы конференции Информационные системы и технологии ИСТ - 2007. Н. Новгород, НГТУ им. Р.А. Алексеева, 2007. с. 225.
93. Савченко В.В., Акатьев Д.Ю., Губочкин И.В. Формирование фонетической базы данных из речевого сигнала на основе информационной теории восприятия речи. // Системы управления и информационные технологии. 2008. 4.1 (34). С. 193-198.
94. Савченко В.В., Губочкин И.В. Фонетический анализ речи методом переменного дерева. // Изв. вузов России. Радиоэлектроника. 2008. Вып. 3. С. 14-20.

95. Савченко В.В., Губочкин И.В. Оптимизация авторегрессионной модели сигналов в задаче автоматического распознавания речи. // Изв. вузов России. Радиоэлектроника. 2008. Вып. 2. С. 26-31.
96. Савченко В.В., Акатьев Д.Ю. Теоретико-информационное обоснование метода обеляющего фильтра в задачах автоматической обработки речи. // Системы управления и информационные технологии. 2008. №1 (31). С. 21-30.
97. Таблицы по математической статистике // П. Мюллер, П. Нойман, Р. Шторм. Пер. с нем. под ред. В.М. Ивановой. - М.: Финансы и статистика, 1982.-278 с.
98. Акатьев Д.Ю., Губочкин И.В. Алгоритм автоматического формирования фонетической базы данных национального языка по конечной группе дикторов. Материалы конференции Информационные системы и технологии ИСТ — 2009. Н. Новгород, НГТУ им. Р.А. Алексеева, 2009. с. 320-321.
99. Савченко В.В., Акатьев Д.Ю., Губочкин И.В., Карпов Н.В., Пономарёв Д.А. Информационная система фонетического анализа слитной речи: Программа для ЭВМ. / Роспатент. Свидетельство о гос. регистрации № 2008615442 по заявке 2008614233 от 15.09.2008.
100. Савченко В.В., Акатьев Д.Ю., Губочкин И.В. Исследование звукового строя национального языка на основе информационной теории восприятия речи. // Вестник Нижегородского университета им. Н.И. Лобачевского. № 3(1), 2010. с. 215 - 222.
101. Рау Ф. Ф., Слезина Н. Ф. Методика обучения произношению в школе глухих.-М.: Просвещение. 1981.
102. Комаров К.В. Методика обучения русскому языку в школе для слабослышащих детей. - М.: Оникс 21 век. 2005.
103. Кукушкина О.И., Королевская Т.К., Зеленская Ю.Б. Информационные технологии в обучении произношению. - М.: Полиграф-Сервис. 2004.
104. Hartis A. Computer-Based Audio-Visual Feedback Using Interactive Visual Displays for Speech Training. - PhD thesis. Department of Computer Science, University of Sheffield, 1999.

105. Brooks S., Fallside F., Gulian E., Hinds P. "Teaching vowel articulation with the computer vowel trainer: Methodology and results", 1981, British Journal of Audiology, vol. 15, pp. 151-163.
106. Савченко В.В. Информационная теория обучения речи // Изв. вузов России. Радиоэлектроника. 2009. Вып. 3. С. 3-12.
107. Савченко В. В., Савченко А. В. Принцип минимального информационного рассогласования в задаче распознавания дискретных объектов // Изв. вузов России. Радиоэлектроника. 2005. Вып. 3. С. 10-18.
108. Пономарев Д.А. Двухэтапный алгоритм обучения речи слабослышащих на основе информационной теории восприятия речи. // Информационные технологии. № 9, 2009. с. 73 - 77.
109. Савченко В.В., Акатьев Д.Ю., Губочкин И.В., Пономарёв Д.А. Автоматизированная система обучения речи глухонемых и слабослышащих: Программа для ЭВМ. / Роспатент. Свидетельство о гос. регистрации № 2009615092 по заявке 2009613925 от 20.07.2009.
110. Савченко В.В., Акатьев Д.Ю., Губочкин И.В. Автоматизированная система обучения речи на основе теоретико-информационного подхода. // Исследовано в России, 1243-1252, 2009. URL: http://zhurnal.ape.relarn.ru/articles/2009/099.pdf.
111. Савченко В.В., Акатьев Д.Ю., Губочкин И.В. Результаты натурных испытаний автоматизированной системы обучения речи слабослышащих на основе информационной теории. // Исследовано в России, 1441-1449, 2009. URL: http://zhurnal.ape.relarn.ru/articles/2009/109.pdf.
112. Акатьев Д.Ю., Губочкин И.В. Информационная система обучения речи глухих и слабослышащих. - Современные проблемы информатизации в моделировании и социальных технологиях: Сб. трудов. Вып. 15 / Под ред. д.т.н., проф., О.Я. Кравца. - Воронеж: «Научная книга», 2010. с. 152 -157.
113. Акатьев Д.Ю., Губочкин И.В. Информационная система обучения речи глухих и слабослышащих на морфологическом уровне. Материалы

конференции Информационные системы и технологии ИСТ — 2010. Н. Новгород, НГТУ им. Р.А. Алексеева, 2010.

114. Макхоул Дж., Рукос С., Гиш Г. Векторное квантование при кодировании речи. // ТИИЭР, т. 73, № 11, 1985. с. 19 -61.
115. Губочкин И.В. Повышение вычислительной эффективности алгоритма распознавания речи на основе метода обеляющего фильтра. // Информационные технологии моделирования и управления. №6, 2010.
116. Винцюк Т.К. Анализ, синтез и интерпретация речевых сигналов. / Киев: Наук, думка, 1987. 264 с.
117. Gray A., Markel J. Distance measures for speech processing. // IEEE Trans. On Acoust., Speech and Lang, processing. Vol. 24 (5), oct. 1976. p. 380 - 291.
118. Губочкин И.В. Применение метода отображения многомерных данных в задаче постановки правильного произношения звуков в составе отдельных слов. // Вестник Новосибирского государственного университета. Серия информационные технологии. 2010, т. 8, вып. 4, с. 61-71.
119. Бондарев, С. Ю., Ситников И.В. Метод биометрической идентификации личности по голосу. С. Ю. Бондарев // Инженерные кадры – Будущее инновационной экономики России. Материалы II Всероссийской студенческой конференции. Йошкар-Ола, 21-25 ноября 2016 г. – Часть 4. Информационные технологии – основа стратегического прорыва в современной промышленности. – Й-Ола: ПГТУ, – 2016. – С. 11-15. URL: https://www.volgatech.net/upload/documents/science/%D0%A7%D0%B0%D1%81%D1%82%D1%8C%204-%D0%98%D0%BD%D0%BD%D0%BE%D0%B2%D0%B0%D1%86%D0%B8%D0%B8.pdf (дата обращения: 30.05.2017)
120. Significance of parametric spectral ratio methods in detection and recognition of whispered speech //Mathur et al. EURASIP Journal on Advances in Signal Processing 2012, 2012:157, – Springer Open Journal, – 2012 – P. 20. URL: https://asp.eurasipjournals.com/content/2012/1/157(дата обращения: 24.11.2016)