**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования**

**«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ЛИНГВИСТИЧЕСКИЙ УНИВЕРСИТЕТ им. Н.А. ДОБРОЛЮБОВА»**

**В.В. Савченко, Д.Ю. Акатьев**

# ИНФОРМАЦИОННАЯ ТЕОРИЯ ВОСПРИЯТИЯ РЕЧИ

## Учебное пособие

**Нижний Новгород**
**2015**

Печатается по решению редакционно-издательского совета ФГБОУ ВПО «НГЛУ».

Направления подготовки: 10.03.01 – *Информационная безопасность*; 45.03.02 – *Лингвистика*.

Дисциплина: Информационная теория восприятия речи.

УДК 81’322(075.8)
ББК 32.97
С 137

Савченко В.В., Акатьев Д.Ю. Информационная теория восприятия речи: Учебное пособие. – Н. Новгород: ФГБОУ ВПО «НГЛУ», 2015. – 92 с.

В учебном пособии изложены основные понятия и определения информационной теории восприятия речи, в которой создается необходимая теоретическая база для дальнейшего изучения математических основ и возможностей современных методов автоматического анализа и распознавания речи в задачах обработки информации.

УП предназначено для самостоятельной работы студентов информационных и лингвистических направлений подготовки в НГЛУ, а также для аспирантов по направлению подготовки 09.06.01 – *Информатика и ВТ* (профиль «Теоретические основы информатики») и специалистов широкого профиля, желающих повысить свою квалификацию в одной из наиболее динамично развивающихся областей современного научного знания.

УДК 81’322(075.8)
ББК 32.97

Авторы: В.В. Савченко, д.т.н., профессор, зав. кафедрой математики и информатики

Д.Ю. Акатьев, к.т.н., профессор кафедры математики и информатики

Рецензент Ю.В. Воронков, к.т.н., доцент, директор отделения очно-заочного и заочного обучения

© ФГБОУ ВПО «НГЛУ», 2015
© Савченко В.В., Акатьев Д.Ю., 2015

## СОДЕРЖАНИЕ

**Предисловие** 4

**Введение** 6

**1. Математические методы и модели речевого сигнала** 9

1.1. Проблема априорной неопредёленности и экстремальные распределения 9

1.2. Модель случайного гауссовского процесса 12

1.3. Задача распознавания образов 13

1.4. Корреляционные свойства СГП 14

**2. Метод обеляющего фильтра** 17

2.1. Задача распознавания речи 17

2.2. Задача распознавания личности по голосу 21

2.3. Пример практического применения 25

2.4. Обсуждение теоретической части 31

**3. Информационная система фонетического анализа речи** 33

3.1. Основные элементы ИТВР 34

3.2. Архитектура информационной системы 35

3.3. Описание программы для ПК 41

3.4. Примеры применения 64

**4. Задания к лабораторному практикуму** 67

**Библиографический список** 68

**Приложение 1** 70

**Приложение 2** 88

## Предисловие

Информационная теория восприятия речи **(**ИТВР) – новейшая теория обработки речевого сигнала на фоне помех в задачах анализа и распознавания речи. Её теоретическая значимость состоит в создании эффективного математического аппарата, нацеленного на преодоление проблемы вариативности устной речи. До настоящего времени указанная проблема получила решение в лабораторных условиях при отсутствии внешних помех. Вторая актуальная проблема в задачах автоматического распознавании речи (АРР) – помехи, в связи с которыми возникают два вида ошибок – ложная тревога (когда фрагмент шума определяется, как фонема) и ошибка пропуска (когда из-за высокого уровня шума распознавание фонемы невозможно).

Проблема, связанная с сопровождением речевого сигнала помехами до настоящего времени не преодолена. ИТВР – единственная на данный момент теория и одновременно учебная дисциплина, нацеленная на решение данной проблемы. Пилотным образцом такого решения может служить новейший метод АРР: фонетического декодирования слов, основанный на правилах слоговой фонетики разговорного русского языка и представляющий собой основной предмет исследований с позиций ИТВР.

Ещё одно важное назначение ИТВР – конструктивное моделирование процесса формирования сигнала в речевом тракте. Принадлежащая ИТВР адаптивная кластерная модель минимальных речевых единиц (элементарного речевого сигнала) является эффективным инструментом для идентификации как речевых образцов, так и личностей дикторов. Отметим, что обе задачи представляют первостепенный интерес с точки зрения разработки и применения современных речевых технологий широкого назначения: от задач анализа до систем голосового управления. Поэтому предлагаемое учебное пособие рассчитано на студентов разных

направлений подготовки и специалистов широкого профиля, желающих повысить свою квалификацию в одной из наиболее динамично развивающихся областей современного знания.

## Введение

Речевая функция является продуктом высшей нервной деятельности человеческого организма и служит непременным условием абстрактного или образного способа мышления. Образ каждого явления в сознании человека сопровождается соответствующей речевой «меткой». По ней данный образ и определяется (опознается) в процессе межличностных речевых коммуникаций. Причём опознавание образов одновременно осуществляется по двум противоположным направлениям. В режиме речеобразования решается обратная задача: подбор последовательности речевых меток под определенный набор образов в сознании человека. А в режиме восприятия речи это прямое опознавание образов, но только по меткам другого человека – собеседника. Именно здесь и возникают основные проблемы с изучением речевого механизма. Первостепенная из них – проблема вариативности речи. Один и тот же образ в восприятии разных людей имеет разные (на слух) речевые метки. В какой степени разные? Как это сказывается на точности восприятия речи? И где предел допустимых различий одноименных меток при речевых коммуникациях? – Всё это ключевые вопросы по теории восприятия речи. Ответы на них на базовом, фонетическом уровне восприятия речи, т. е. на уровне звукового строя национального языка, и составляют главное содержание **ИТВР**. При этом получены следующие фундаментальные результаты:

1. На основе вариационного подхода и натуральной параметризации данных поставлена и решена задача восстановления неизвестного закона распределения элементарного речевого сигнала по первым двум его выборочным моментам. Показано, что оптимальным в теоретико-информационном смысле в расчёте на задачи распознавания образов является многомерный гауссов закон. Рассмотрен пример его практического применения в задаче автоматического распознавания речи как актуальной разновидности задачи статистической классификации.

Дано строгое теоретическое обоснование критерия минимума информационного рассогласования (МИР) как оптимального в байесовском смысле.

2. На основе гауссовой авторегрессионной модели речевого сигнала с нормировкой средней мощности порождающего воздушного потока на входе «акустической трубы» речевого механизма человека предложен новый подход к реализации критерия МИР: по методу адаптивного обеляющего фильтра (МОФ) строго доказана его асимптотическая оптимальность в байесовском смысле для задач статистической классификации «с учителем».

3. На основе МОФ впервые показано: теоретически и экспериментально, что гауссова аппроксимация речевого сигнала гарантирует достаточно высокую вероятность безошибочного распознавания даже наиболее проблемных с точки зрения своей различимости речевых единиц. Указанный эффект особенно значителен в условиях конечных (малых) интервалов наблюдения. Большинство задач в области автоматической обработки речи, таких как фонетический анализ разговорной (устной) речи (ФАР), относится именно к такому случаю.

4. Задача ФАР впервые строго сформулирована как разновидность статистической классификации «без учителя», а её асимптотически минимаксное решение впервые сведено к критерию МИР и МОФ при реализации последнего на основе решетчатой структуры адаптивного обеляющего фильтра. Алгоритмическую основу нового подхода к ФАР образует процедура группирования последовательности элементарных речевых единиц (ЭРЕ), полученной в результате линейного членения речевого сигнала в несколько множеств типа кластеров, организованных по принципу МИР вокруг своих информационных центров-эталонов в метрике Кульбака-Лейблера.

5. По результатам решения задачи ФАР впервые строго сформулировано понятие информационного центра-эталона (ИЦЭ) множества одноименных (подобных) ЭРЕ, а вслед за ним определено и ключевое для новой теории восприятия речи понятие фонемы как множества подобных ЭРЕ, объединенных по критерию МИР в общий кластер вокруг соответствующего ИЦЭ.

По сути, данным теоретико-информационным определением ЭРЕ и была создана, в основном, ИТВР как актуальное направление фундаментальных исследований в области акустической теории речеобразования, уходящей своими корнями в классические работы Ch.G. Kratzenstein (1779), H. Helmholtz (1870) и других исследователей. Благодаря ИТВР появляются новые возможности не только эффективно моделировать человеческий механизм восприятия речи, но и усиливать его качественные характеристики, в частности, индивидуальные особенности звукового строя разговорного языка разных дикторов. Поэтому предложенная теория может найти практическое применение при решении широкого круга актуальных задач в области как теоретической, так и прикладной информатики. Её базовые положения рассмотрены в настоящем учебном пособии.

# 1. Математические методы и модели речевого сигнала

## 1.1. Проблема априорной неопределённости и экстремальные распределения

До недавнего времени наиболее распространённым подходом в области речевых технологий являлся теоретико-вероятностный или статистический подход, реализованный в большой группе методов, разработанных на основе скрытой марковской модели (СММ) речевого сигнала, или СММ-методах. Указанная модель отталкивается, в свою очередь, от понятия экстремального закона распределения. **Экстремальным** называется такой закон, который при определённых условиях или ограничениях на плотность вероятности $f^*(x)$ случайной величины $X$ обеспечивает максимум (экстремум) её энтропии. Именно экстремальному распределению с максимальной энтропией отдают свое предпочтение исследователи в условиях полной **априорной неопределенности в отношении** закона распределения случайной величины. Обоснованием сказанного может служить следующая теорема.

**Теорема о концентрации энтропии.** *Из всех гипотетических распределений закон с максимальной энтропией* $f^*(x)$ *реализуется при многократных наблюдениях наибольшим числом способов*.

Применительно к ИТВР принцип максимума энтропии служит способом преодоления априорной неопределенности в отношении закона распределения речевого сигнала $X(t)$ на интервале времени $t \in (t_i; t_i + \tau_0)$ длительностью в один фрейм $\tau_0$ = 5-20 мс.

Рассмотрим следующие примеры экстремальных распределений.

**Задача 1.**

Пусть $X$ – случайная величина с распределением в области $x \in [a,b]$. Найдите экстремальный закон распределения для неё.

**Решение.**

Рассмотрим дифференциальную энтропию

$$H(x) \overset{\Delta}{=} -\int_{-\infty}^{+\infty} f(x) \ln f(x) dx.$$

Потребуем от плотности экстремального закона её максимизации:

$$H(f^*(x)) = \max$$

Единственным ограничением в этой задачи является условие нормировки плотности вида

$$\int_{-\infty}^{+\infty} f(x) dx = 1.$$

Задача после этого формулируется как **вариационная**. Воспользуемся для решения стандартным методом Лагранжа (метод неопределенных множителей). Сначала вычислим функционал («функцию от функции») вида

$$L\{f(x)\} = H(X) + \lambda_1 [\int_a^b f(x) dx - 1].$$

Затем возьмем от него вариацию по переменной *f*:

$$\nabla L\{f(x)\} = -1 - \ln f(x) + \lambda_1 = 0.$$

Приравняв вариации к 0, получим оптимизационное уравнение

$$\ln f(x) - \lambda_1 = 1.$$

Здесь $\boldsymbol{\lambda}_1$ – неопределенный множитель Лагранжа (const), т. е. имеем равенство

$f^*(x) = c = \text{const}$.

Значение константы находится из уравнения-ограничения нормировки

$$\int_a^b c dx = (b-a)c = 1, \quad c = \frac{1}{b-a}.$$

В результате мы получили равномерный закон в пределах заданного интервала значений – в качестве искомого экстремального.

Вывод: В условиях ограниченной области определения [$a$, $b$] случайной величины $X$ именно равномерный закон выполняет роль закона экстремального как наилучшей аппроксимации неизвестного истинного распределения.

**Задача 2.**

Усложним формулировку нашей задачи

$$H(x) \overset{\Delta}{=} -\int_{-\infty}^{+\infty} f(x)\ln f(x)dx = \max,$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1,$$

вводя дополнительное ограничение на дисперсию случайной величины

$\mathrm{D}(X) = \sigma^2 = \mathrm{const}$

в предположении, что её математическое ожидание $\mathrm{M}(X) = m$ нам известно. Например, во многих случаях априори имеем $m = 0$ (это относится к тепловому шуму на выходе канала связи, но также и к речевому сигналу в общем случае).

**Решение.**

По аналогии с предыдущей задачей запишем функционал Лагранжа вида

$$L\{f(x)\} = H(X) + \lambda_1[\int_{-\infty}^{\infty}(x-m)^2 f(x)dx - \sigma^2] + \lambda_2[\int_{-\infty}^{\infty} f(x)dx - 1].$$

Возьмем от него вариацию по той же переменной $x$. Получим следующее оптимизационное уравнение:

$$\nabla L\{f(x)\} = -1 - \ln f(x) + \lambda_1(x-m)^2 + \lambda_2 = 0.$$

Решая его, получим

$$f^*(x) = c\exp(-\lambda_1(x-m)^2) = 0,$$

где $c$, $\lambda_1$ – две константы. Это нормальный закон распределения общего вида

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-m)^2\}.$$

Поэтому именно гауссовский закон используется в современных исследованиях и разработках в области речевых технологий как наиболее подходящая аппроксимация для распределения речевого сигнала. На основании сказанного далее обсуждается универсальная статистическая модель речевого сигнала.

## 1.2. Модель случайного гауссовского процесса

Случайный гауссовский процесс (СГП) широко используется для математического моделирования (аппроксимации) речевого сигнала. СГП стационарного типа $X(t)$ задается следующим выражением для своей многомерной плотности вероятности:

$$f_n(x_1 \ldots x_n) \stackrel{\Delta}{=} f_n(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det K_{xx}}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{m})^T K_{xx}^{-1}(\vec{x}-\vec{m})\right\}.$$

Здесь $K_{\tilde{o}\tilde{o}} - (n \times n) -$ матрица коэффициентов корреляции $k_{ij}$, $i, j \le n.$, или автокорреляционная матрица (АКМ) СГП. Многомерная плотность – это исчерпывающая математическая формулировка СГП, причём формулировка параметрического типа. Иными словами, задать СГП строго математически – это значит, задать его АКМ. Отметим, что для стационарного СГП $k_{ij} = k_{ji}$ АКМ – это строго симметричная матрица с дисперсиями $\sigma^2_{\tilde{o}}$ на своей главной диагонали. В результате, в рамках гауссовской модели обработка речевого сигнала сводится, по существу, к корреляционной или к вычислению оценок автокорреляционной функции (АКФ) случайного процесса вида:

$$k(\tau) = M\{X(T) * X(T+\tau)\}.$$

При этом нетрудно убедиться, что $k(0) = \sigma^2_{\tilde{o}}$.

Основное свойство АКФ – симметричность относительно начала координат. Значения АКФ в дискретных точках определяет соответствующие элементы АКМ случайного процесса $X(t)$.

### 1.3. Задача распознавания образов

Наиболее общая формулировка большинства задач автоматической обработки и распознавания речи может быть дана в терминах теории распознавания образов: требуется отнести предъявляемый объект наблюдения $X$ (в нашем случае – выборка из речи) к одному из $R > 1$ классов (личностей или отдельных слов) заранее точно не определенных. Каждый класс характеризуется тем свойством, что принадлежащие ему объекты (образцы речи данной личности) обладают некоей общностью, сходством. То общее, что объединяет объекты в класс, и называют *образом*. Иными словами, каждую конкретную личность мы априори задаем в виде её вполне определенного образа или набора (множества) устойчивых признаков $\mathbf{P}_r, r = \overline{1, R}$. В таком случае решение рассматриваемой задачи сводится к установлению отношения эквивалентности $\mathbf{P}_X = \mathbf{P}_\nu, \quad \nu \le R$, между соответствующим набором признаков объекта наблюдения $X$ и одним из его $R$ ($\nu$-м) образом из нашей базы априорных данных. Проблема состоит в том, что каждой конкретной личности присуща известная вариативность, т. е. изменчивость ее голоса от одного произношения к другому, которая носит к тому же неопределённый, случайный характер. Обычно решение данной проблемы связывают [1; 2] со статистическим подходом, когда в роли каждого образа $\mathbf{P}_r, r = \overline{1, R}$ выступает соответствующий закон распределения многомерной выборки наблюдений из некоторой гипотетической генеральной совокупности $\{\mathbf{R}^n, \mathbf{P}_r\}$. Здесь $n$ –

размерность выборки, равная числу имеющихся отсчетов речевого сигнала, а $\mathbf{R}^n$ – многомерное евклидово пространство. Задача установления отношения эквивалентности общего вида переходит в таком случае в задачу проверки статистических гипотез о неизвестном законе распределения. Но здесь возникает новое препятствие, а именно: проблема встречных гипотез в отношении вида каждого из $R$ альтернативных законов. Радикальное средство для её преодоления – это восстановление неизвестного закона $\mathbf{P}_r, r = \overline{1, R}$, в процессе предварительного обучения системы автоматического распознавания образов. Указанное обучение осуществляется в процессе показа ряда образцов человеческого голоса $X_r = \{\vec{x}_l^{(r)}\}, r = \overline{1, R}$, принадлежность которых к тому или иному классу (образу) точно известна. Наиболее строго данная проблема решается на основе гауссовской параметризации речевого сигнала и при условии его центрирования, когда выполняется векторное равенство $\vec{m} = \vec{0}$. Подобная задача подробно рассматривается в следующей главе.

## 1.4. Корреляционные свойства СГП

Корреляционные свойства СГП однозначно определяются его АКМ. Обозначим её применительно к некоторому процессу $X(t)$ как $K_{xx} = \left\| k_{ij} \right\|, \ i, j \le n.$ Здесь каждый элемент матрицы определяет соответствующий корреляционный момент (коэффициент) для фиксированной пары $i$-го и $j$-го отсчетов $x(i)$ и $x(j)$ соответственно: $k_{ij} = M\{\overset{o}{x}(i) * \overset{o}{x}(j)\},$ где $\overset{o}{x}$ – центрированный СП. Данное определение автокорреляционной матрицы может быть сведено к векторно-матричной форме $K_{xx} = M(\vec{x}\vec{x}^T), \ \vec{x} = col\{x(i)\}_1^n.$ Матрица определяет зависимость каждого

элемента к каждому, а симметрична она потому, что $i$ и $j$ дважды повторяются.

Таким образом, корреляционная связь, в отличие от других характеристик между детерминированными или случайными явлениями представляет собой особый вид статистической связи (взаимосвязи), проявляемый исключительно на множестве наблюдений в отличие, например, от детерминированной формы связи. Результат – корреляционный анализ находит широкое применение при исследовании, прежде всего, случайных явлений со слабо предсказуемым поведением в будущем. Учитывая, что именно такие явления более распространены в задачах экономики и управления, можно утверждать, что именно корреляционный анализ образует основу большинства экономических исследований, включает исследования и разработки в области информационной безопасности.

*Пример.*

Пусть $n = 3$, тогда отсчёты $x(t) = x(1), x(2), x(3)$ образуют вектор – столбец $\vec{x} = \{x_1, x_2, x_3\}^T$. Определим по нему матрицу

$$R_{xx} = \begin{bmatrix} x_1 & 0 & 0 \\ x_2 & 0 & 0 \\ x_3 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} x_1 x_2 x_3 \\ 0\ 0\ 0 \\ 0\ 0\ 0 \end{bmatrix},$$

дополняя в векторе $\vec{x} = \{x_1, x_2, x_3\}^T$ недостающие до квадратной матрицы элементы нулями. Возьмем от этой матрицы математическое ожидание $\mathbf{M}(\cdot)$ и получим АКМ вида

$$\mathbf{M}(R_{xx}) = \begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{bmatrix} = K_{xx}.$$

Отметим симметричность данной матрицы: $k_{ij} = k_{ji}\ \forall i, j \le n,$ в общем случае. Более того, в случае стационарного процесса значения становятся зависимыми только от временного интервала $\Delta = j\text{-}i$ между ними, а именно:

$$K_{xx} = \begin{bmatrix} \hat{e}(0) & \hat{e}(1) & \hat{e}(2) \\ \hat{e}(-1) & \hat{e}(0) & \hat{e}(1) \\ \hat{e}(-2) & \hat{e}(-1) & \hat{e}(0) \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \hat{e}(1) & \hat{e}(2) \\ \hat{e}(1) & \sigma_x^2 & \hat{e}(1) \\ \hat{e}(2) & \hat{e}(1) & \sigma_x^2 \end{bmatrix}.$$

И, наконец, получим АКМ теплицевой структуры:

$$K_{xx} = \sigma_x^2 \cdot \begin{bmatrix} 1 & \rho(1) & \rho(2) \\ \rho(1) & 1 & \rho(1) \\ \rho(2) & \rho(1) & 1 \end{bmatrix},$$

где $\rho(i) = \dfrac{\hat{e}(i)}{\sigma_x^2}$ – нормированный коэффициент корреляции в зависимости от сдвига $i$ между двумя сечениями СГП в дискретном времени.

## 2. Метод обеляющего фильтра

### 2.1. Задача распознавания речи

Проще всего поставленная задача

$$\mathrm{P}_x = \mathrm{P}_r, \quad r = 1,2,\ldots,R \tag{1}$$

решается в параметрическом варианте, когда каждое распределение $\mathrm{P}_r$, $r \in \overline{1, R}$ берется нами из некоторого параметрического семейства гауссовских законов распределения $\mathrm{P}_r = N(\mathrm{K}_r)$, определённых на множестве допустимых значений всех элементов $(n \times n)$-матрицы автокорреляций $\mathrm{K}_r$, $r = \overline{1, R}$ (в данном случае предполагается, что все используемые сигналы заранее центрированы). Восстановление закона $N(\mathrm{K}_r)$ по обучающей выборке $X_r$ сводится к элементарной процедуре статистического оценивания его неизвестной автокорреляционной матрицы (АКМ) по формуле

$$\mathrm{K}r = \frac{1}{L} \sum_{l=1}^{L} \vec{x}_l \vec{x}_l^T \Big|_{\vec{x}_l = \vec{x}_l^r}, \quad r = \overline{1, R}, \tag{2}$$

где $L$ – число независимых образцов речи в пределах $r$-го распределения $\mathrm{P}_r$, а символом "$T$" обозначена операция транспонирования векторов ($\vec{x}_l^{(r)}$ – $n$-вектор-столбец с координатами из $\mathrm{R}^n$). Задача (1) после этого формулируется как проверка $R$ простых гипотез о неизвестном законе распределения:

$$H_r : \mathrm{P}_X = N(\mathrm{K}r), \quad r = \overline{1, R}. \tag{3}$$

Это стандартная задача статистической классификации. Её решение обычно основывается на критерии максимального правдоподобия. Применительно к нашей модели речевых сигналов в виде $L$ независимых отрезков (массивов), длиной $n$ отсчетов каждый, такая задача подробно рассматривалась в работе [1]. Было показано, что оптимальный алгоритм

принятия решения по выборке $X = \{\vec{x}_l\}, l = \overline{1, L}$ основывается на достаточной статистике общего вида

$$\lambda_r(X) = \mathrm{tr}\,[\mathrm{K}_X \mathrm{K}_r^{-1}] + \ln|\mathrm{K}_r|, \quad r = \overline{1, R}, \tag{4}$$

где $\mathrm{K}_X$ – выборочная оценка АКМ наблюдений по формуле (2); символами $\mathrm{tr}(\cdot)$ и $|\cdot|$ обозначены соответственно след и определитель квадратных $(n \times n)$-матриц. Решение принимается в пользу гипотезы $H_\nu$ из (1), если соответствующая ей $\nu$-я статистика (4) принимает минимальное значение из $R$ её рассматриваемых альтернатив.

Указанный алгоритм в асимптотическом варианте (при $n \to \infty$) для несингулярного случая, когда $\lim|\mathrm{K}_r| \neq 0$, может быть переписан следующим образом [4]:

$$\lambda_r(X) = \frac{\sigma_r^2(X)}{\sigma_r^2} + \ln \sigma_r^2, r = \overline{1, R} \tag{5}$$

$$\sigma_r^2(X) = \frac{1}{L} \sum_{l=1}^{L} \left[ y_l^{(r)}(X) \right]^2;$$

$$y_l^{(r)}(X) = \vec{a}_r^T \vec{x}_l; \tag{5'}$$

$$\vec{a}_r = \sigma_r^2 K_r^{-1} \vec{\Gamma};$$

$$\sigma_r^2 = \left[ \vec{\Gamma}^T K_r^{-1} \vec{\Gamma} \right]^{-1},$$

где $\vec{\Gamma} = col_n(1, 0, \ldots, 0)$ – $n$-вектор-столбец, составленный из нулей, за исключением единицы на первой позиции. Решение здесь принимается в пользу гипотезы $H_\nu$ при условии минимизации взвешенной с коэффициентом $1/\sigma_\nu^2$ и смещённой на $\ln \sigma_\nu^2$ величины выборочной

дисперсии $\sigma_{\nu}^{2}(X)$ отклика на сигнал $X$ декоррелятора $\nu$-го канала (5'). Структура такого декоррелятора однозначно определяется вектором коэффициентов линейной авторегрессии (АР-коэффициентов) $\vec{a}_{\nu}, \nu \le R$. Это стандартная формулировка *метода обеляющего фильтра* (МОФ) в задачах распознавания образов [4].

Главная идея метода состоит в существенной (в десятки раз) редукции или сжатии данных за счет того, что в базе априорных данных хранятся не сами отрезки речи длиной $nL = 10^2 \ldots 10^5$ отсчётов каждый, а их образы в виде набора из $R$ векторов АР-коэффициентов (их также часто называют коэффициентами линейного предсказания – КЛП), размерность которых $M = 10 \ldots 30 << nL$ в реальных условиях ограничивается конечной степенью сложности спектрального состава человеческого голоса. Причём, в отличие от известных алгоритмов автоматического распознавания речи на основе коэффициентов линейного предсказания или авторегрессии, в рассматриваемом МОФ применяется принципиально иной критерий для оценивания рассогласования между различными речевыми образами, уходящий своими корнями в теоретико-информационный подход и информационную метрику Кульбака-Лейблера. В работе [2] доказывается следующее положение.

*Утверждение 1*. Выражение для решающей статистики (5) с точностью до константы $(\ln|K_X^{-1}| - n)$ и весового коэффициента 2,0 воспроизводит собой величину информационного рассогласования (ИР) упорядоченной пары гауссовских распределений $P_X$ и $P_r$ по Кульбаку-Лейблеру

$$I_n[P_X | P_r] = 0{,}5[\mathrm{tr}[K_X K_r^{-1}] + \ln|K_X^{-1} K_r| - n]. \qquad (6)$$

Таким образом, оптимальная решающая статистика (4) реализует в своей основе критерий *минимума информационного рассогласования*

(МИР). В таком случае алгоритм распознавания образов в формулировке (5) можно рассматривать как реализацию критерия МИР в терминах МОФ.

*Следствие 1*. Удельная величина (на один отсчет данных) при $n \to \infty$ информационного рассогласования (6) между $\nu$-м и $r$-м образцами речевого сигнала с точностью до константы (-1) и множителя 0,5 равна

$$\gamma_{\nu,r} = \frac{\sigma_r^2(x_\nu)}{\sigma_r^2} + \ln \frac{\sigma_r^2}{\sigma_\nu^2} \ . \qquad (7)$$

*Утверждение 2*. Вне зависимости от вида речевых сигналов удельная величина (7) отвечает соотношению

$$\gamma_{\nu,r} \geq 1$$

с равенством лишь при условии идентичности их образов-распределений $P_\nu = P_r = N(K_r)$ и обладает, поэтому, всеми качествами информационной метрики Кульбака-Лейблера.

*Доказательство* данного утверждения основывается на утверждении 1 и известных свойствах метрики $I_n[P_\nu | P_r]$.

*Следствие 2*. Понятию полностью эквивалентных образов $\nu$-го и $r$-го речевых сигналов соответствует единичное значение информационного рассогласования (ИР) по формуле (7); чем больше эта метрика отличается от единицы, тем больше различия между рассматриваемыми сигналами.

При этом следует особо оговориться, что величина (7) не является метрикой (расстоянием) в строгом смысле, т. к. она имеет несимметричный вид относительно пары распределений $P_\nu$ и $P_r$. Может быть поэтому в своей основополагающей работе сам С. Кульбак предложил оценивать расстояние между распределениями по средней величине информационного рассогласования $(\gamma_{\nu,r} + \gamma_{r,\nu})/2$. Но в нашей

задаче эта деталь не имеет особого значения. Важно понять значение метрики (7) в контексте следствия 2: решая задачу распознавания образов по МОФ (2), (4), мы получили в форме решающей статистики (5) достаточно полную информацию о надёжности нашего результата. Чем сильнее статистика (5) при $r = \nu$ устремляется к своему минимуму при $r = \nu$, тем более надежным следует считать решение в пользу гипотезы $H_\nu$. И, наоборот, при относительной близости по величине всех $R$ статистик (5) любой конкретный результат для данного речевого сигнала $X$ должен вызывать у нас определенные сомнения. По сути, здесь мы достигаем возможности одновременного распознавания и оценивания образцов речи. Именно в этом актуальном направлении исследований и находятся, на наш взгляд, основные преимущества МОФ. Для подтверждения и развития сделанных выводов рассмотрим еще одну актуальную задачу.

### 2.2. Задача распознавания личности по голосу

Пусть, как и ранее, $X = \{\vec{x}_l\}$ – многомерная повторная выборка из речевого сигнала от некоторого неизвестного нам точно источника (личности). А $\mathbf{P}_0 = N(\mathbf{K}_0)$ – нормальный закон распределения, заданный своей $(n \times n)$-матрицей автокорреляций $\mathbf{K}_0$ в роли статистического образа известной личности. Требуется ответить на вопрос: принадлежит ли выборка (голос) $X$ данной личности $\mathbf{P}_0$? Это еще одна модификация задачи распознавания образов для случая дихотомии, т. е. бинарного («да» - «нет») множества принимаемых решений. Задача в такой постановке была подробно изучена в работе [3] в терминах классической задачи «о разладке»: простая гипотеза в отношении вида АКМ $n$-мерного гауссовского распределения

$$H_1 : \mathbf{K}_X = \mathbf{K}_0$$

проверяется против сложной альтернативы

$$H_1 : \mathbf{K}_X \neq \mathbf{K}_0$$

на некотором фиксированном уровне значимости принимаемого решения

$$P(X \in W \mid H_0) = \alpha_0 = const \qquad (8)$$

Здесь критическая область $W$ в пределах выборочного пространства $\mathbf{R}^n$ зависит от критерия решения поставленной задачи. Следуя критерию отношения правдоподобия из работ [4], будем иметь

$$W : \lambda(X) = \mathrm{tr}(\mathbf{K}_X \mathbf{K}_0^{-1}) + \ln|\mathbf{K}_X^{-1} \mathbf{K}_0| > \lambda_0,$$

где $\lambda_0$ – пороговый уровень, в который пересчитывается уровень значимости $\alpha_0$. Полученный результат по своей сути повторяет выражение для *r*-й решающей статистики (5), но только с еще более явной своей привязкой к величине ИР по Кульбаку-Лейблеру (6). И это неудивительно: задача обнаружения разладки предполагает обязательным предварительное оценивание степени рассогласования статистических свойств анализируемой выборки наблюдений $X$ относительно исходного распределения $\mathbf{P}_0$ перед принятием любого обоснованного решения. Оптимальное решающее правило формулируется в таком случае следующим образом [1]: принимается решение в пользу гипотезы $H_1$ (то есть гипотеза $H_0$ о принадлежности голоса данной личности отвергается), если выполняется условие

$$\lambda(X) = \frac{\sigma_0^2(X)}{\sigma_0^2} + \ln\frac{\sigma_0^2}{\sigma_x^2} > \lambda_0, \qquad (9)$$

где величины $\sigma_0^2(X)$ и $\sigma_0^2$ определяются по аналогии с выражением (5) как выборочные дисперсии откликов единственного ($R=1$) обеляющего фильтра (5') (настроен по сигналу $X_0$) на выборку $X$ и обучающий сигнал $X_r$ при $r=0$ соответственно, а $\sigma_x^2$ – минимальное собственное число статистической оценки ($n \times n$)-матрицы автокорреляций $\mathrm{K}_{XX}$, которая

рассчитывается по формуле выборочного смешанного момента второго порядка (2). Строго говоря, по определению [4] при $n \to \infty$ имеем равенство $\sigma_x^2 = \left[\vec{\Gamma}^T \, \mathrm{K}_X^{-1} \vec{\Gamma}\right]^{-1}$. Однако здесь возникает известная [2] проблема плохой обусловленности выборочной оценки матрицы $\mathrm{K}_X$, вследствие которой величина $\sigma_x^2$ во многих случаях, особенно в условиях малых выборок наблюдений, вычисляется с недопустимо большой погрешностью. А это объективно может приводить к дополнительным ошибкам в нашем асимптотически оптимальном решении по правилу (9). Выход из такой ситуации видится в использовании следующего приема. Упростим формулировку (9), заменяя в ней величину $\sigma_x^2$ на выборочную дисперсию $\sigma_0^2(X) \geq \sigma_x^2$ в роли ее оценки «сверху», переходя, тем самым, к оценке «снизу» для величины решающей статистики вида

$$\hat{\lambda}(X) = \frac{\sigma^2(X)}{\sigma_0^2} - \ln \frac{\sigma^2(X)}{\sigma_0^2}. \qquad (10)$$

*Утверждение 3.* Статистика (10) имеет единственный минимум, равный 1, при равенстве $\sigma_0^2(X) = \sigma_0^2$, т.е. $\forall X \in \mathrm{R}^n: \hat{\lambda}(X) \geq 1$.

*Доказательство.* В обозначении $\dfrac{\sigma_0^2(X)}{\sigma_0^2} = y$ будем иметь функцию $\hat{\lambda}(X) = \hat{\lambda}(y) = y - \ln y$. Её первая производная по переменной $y$ равна $\hat{\lambda}'(y) = 1 - \dfrac{1}{y}$. Приравнивая её к нулю, получаем единственную стационарную точку $y^* = 1$, причем со свойством минимума, т. к. $\hat{\lambda}''(y^*) = y^{-2} > 0 \; \forall y > 0$.

*Утверждение 4.* Решающее правило (9) с модифицированным видом решающей статистики (10) характеризуется асимптотически оптимальным уровнем значимости (8) при неизменном пороговом уровне $\lambda_0 = const$.

*Доказательство* последнего положения автоматически вытекает из того, что при справедливости гипотезы $H_0$ рассматриваемая величина

$\hat{\lambda}(X)$ является состоятельной оценкой оптимальной решающей статистики (9) в силу очевидной условной сходимости выборочных дисперсий $\sigma_0^2(X)$ и $\sigma_x^2$ к одной и той же дисперсии $\sigma_0^2$ порождающего шума в авторегрессионной модели наших наблюдений [4], т. е. $\sigma_0^2(X) \xrightarrow{п.н.} \sigma_0^2 = \sigma_x^2 |_{H_0}$. Здесь сокращение «п. н.» обозначает сходимость «почти наверное» или с вероятностью 1.

Таким образом, в форме выражения (10) мы получили статистическую оценку величины ИР по Кульбаку-Лейблеру в задаче распознавания личности по голосу. Причем точность такой оценки монотонно возрастает по мере приближения ее величины к минимальному единичному уровню (признак идеального соответствия предъявляемого голоса данной личности). Иными словами, во всех тех ситуациях, когда статистика (10) с большой вероятностью не превышает установленный в (9) порог $\lambda_0 > 1$, т. е. принимается решение в пользу гипотезы $H_0$, мы располагаем дополнительной информацией о степени надежности нашего решения. Важность такой информации может быть проиллюстрирована на целом ряде прикладных задач. Например, в задачах судебной экспертизы любое решение оформляется с указанием степени его надежности. Для данной цели можно предложить показатель вида $100/\hat{\lambda}(X)$,%. Значению $\hat{\lambda}(X) = 1$ здесь соответствует 100%-надежное решение (разумеется, это идеал). Другой пример: задача распознавания личности по голосу в многоальтернативной постановке, когда число образов в нашей базе данных $R > 1$. Нетрудно увидеть, что такая задача путем последовательных упрощений сводится к рассмотренной выше дихотомии. Оптимальному решению в пользу конкретной ν-й личности ($\nu < R$) будет отвечать тот же критерий МИР вида $H_\nu : \min_r \hat{\lambda}_r(X)$, где все статистики $\hat{\lambda}_r(X), r = \overline{1, R}$,

определяются согласно выражению (10), но только каждая для своего (одноименного) обеляющего фильтра (5’).

### 2.3. Пример практического применения

Большинство современных методов АРР основывается на процедуре предварительного сегментирования сигналов на короткие отрезки данных. В результате задача сводится к поэтапному распознаванию речи. На первом, базовом этапе распознаются элементарные речевые единицы (ЭРЕ) типа отдельных фонем. На втором, заключительном этапе распознаются слова, фразы и целые тексты как соответствующим образом структурированные последовательности разных ЭРЕ. Таким образом, задача распознавания ЭРЕ является ключевой при АРР. Рассмотрим её подробнее.

В общем случае распознавание ЭРЕ по МОФ сводится к следующему правилу: сначала производится параллельная обработка анализируемого сигнала $X$ в системе из $R$ настроенных на соответствующие эталоны $X_r$ ,$r=\overline{1,R}$ обеляющих фильтров. Затем их отклики сравниваются между собой, и принимается решение в пользу сигнала $X_\nu$ ,$\nu \le R$ с минимальной выборочной дисперсией отклика. Из работы [2] известно, что при общих допущениях о свойствах ЭРЕ в условиях гауссовой аппроксимации их распределений выражение для оптимальной решающей статистики МИР сводится к следующему виду:

$$\rho_{X,r}=\frac{1}{F}\sum_{f=1}^{F}\frac{\left|1+\sum_{m=1}^{p}a_r(m)exp(-j\pi mf/F)\right|^2}{\left|1+\sum_{m=1}^{p}a_X(m)exp(-j\pi mf/F)\right|^2}-1\ge 0. \qquad (11)$$

Это стандартная формулировка выборочной оценки удельной (на один отсчет данных) величины информационного рассогласования между сигналом $X$ на входе и $r$-м сигналом из $R$-словаря в частотной области. Здесь $\{a_r(m)\}$, $\{a_x(m)\}$ – это $p$-векторы коэффициентов линейной среднеквадратической авторегрессии сигналов $X_r$ и $X$ соответственно, $f$ – дискретная частота, а $F$ – верхняя граница частотного диапазона или используемого для передачи речевых сигналов канала связи.

В работе [4] также показано, что эффективность МОФ зависит, прежде всего, от матрицы $R \times R$ значений ВИР $\|\rho_{\nu,r}\|$ между рассматриваемыми сигналами в пределах обучающего словаря. При этом вероятность перепутывания ν-го и $r$-го сигналов определяется в первом приближении выражением

$$\alpha_{\nu,r} = [1 - \Phi_{L,L}(1 + \rho_{\nu,r})],$$

где $\Phi_{L,L}(\cdot)$ – интегральная функция $F$-распределения Фишера с $(L,L)$ степенями свободы, значения которой табулированы. Здесь $L$ – это объем выборки на интервалах квазистационарности (10…20 $mc$) речевого сигнала. Чем больше ВИР между разными сигналами, тем меньше вероятность соответствующих ошибок их распознавания.

Пары ЭРЕ с номерами (ν, $r$), для которых величина $\rho_{\nu,r}$ минимальна, будем называть в дальнейшем проблемными парами – по признаку максимальной вероятности их перепутывания. Именно на таких парах сигналов должно быть сосредоточено наше внимание при обосновании МОФ в задачах АРР. Указанная логика была реализована в дальнейшем при экспериментальных исследованиях МОФ.

Для исследований были выбраны двадцать наиболее распространенных фонем русского языка: |а|, |о|, |у|, |э|, |ш|, |щ|, |р|, |с|, |в|, |з|, |ж|, |и|, |л|, |л’|, |ф|, |х|, |ч|, |е|, |ы|, |м|. Все они последовательно во

времени, многократно (в разных реализациях) проговаривались группой из трёх дикторов, мужчинами разного возраста, в режиме продолжительного (до 1 *с*), достаточно информативного звучания. Полученные сигналы с помощью стандартной аппаратуры и встроенного АЦП (частота дискретизации 8 кГц) были записаны в память ПК в виде соответствующих звуковых файлов. При этом объём выборки по каждой ЭРЕ был установлен заведомо большим: порядка нескольких тысяч единиц. По каждому диктору сначала был составлен отдельный экспериментальный (рабочий) словарь (ЭС) из двадцати эталонных реализаций ЭРЕ. Далее, по каждому сигналу ЭРЕ из ЭС был рассчитан вектор АР-коэффициентов – сначала заведомо невысокого порядка $p = 12$. Для этого применялась высокоскоростная рекуррентная процедура Берга-Левинсона.

По этим векторам были настроены на параллельную обработку входного сигнала $X$ двадцать линейных (трансверсальных) обеляющих фильтров – все одинакового порядка $p$. С их помощью при равенстве $X = X_\nu$, $\nu \le R$, согласно выражению (11), была рассчитана матрица ВИР $\left\| \rho_{\nu,r} \right\|$ размера 20x20. Её фрагмент для первого диктора представлен в виде таблицы 1.

Таблица 1

| | **а** | **ш** | **щ** | **в** | **з** | **л’** | **ф** | **х** | **ч** |
|---|---|---|---|---|---|---|---|---|---|
| **щ** | 95,72 | 4,281 | 0 | 22,73 | 6,161 | 18,9 | 9,157 | 146,3 | **0,126** |
| **р** | 2,908 | 3,787 | 28,88 | 0,839 | 0,812 | 5,67 | 0,527 | 3,121 | 32,57 |
| **с** | 21,09 | 0,448 | 2,815 | 30,58 | 5,916 | 38,3 | 6,888 | 14,28 | 3,002 |
| **в** | 1,485 | 4,832 | 27,23 | 0 | 0,745 | 3,34 | 0,308 | 4,029 | 33,86 |
| **з** | 8,092 | 1,816 | 8,948 | 1,198 | 0 | 2,53 | **0,265** | 5,852 | 9,549 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **ж** | 65,76 | 0,511 | 2,737 | 11,81 | 1,778 | 8,65 | 4,788 | 45,05 | 3,556 |
| **и** | 49,00 | 6,165 | 16,49 | 6,537 | 4,039 | 6,97 | 4,258 | 123,3 | 26,24 |
| **л** | 1,312 | 10,85 | 40,62 | 0,977 | 2,598 | 15,0 | 1,716 | 8,584 | 46,59 |
| **л’** | 26,60 | 3,664 | 15,65 | 2,978 | 1,451 | 0 | 1,719 | 58,45 | 21,64 |
| **ф** | 2,610 | 2,685 | 11,95 | 0,378 | **0,139** | 2,66 | 0 | 5,790 | 13,88 |
| **х** | 2,177 | 6,586 | 31,44 | 1,017 | 0,861 | 5,28 | 0,597 | 0 | 31,13 |
| **ч** | 106,7 | 3,491 | **0,207** | 52,21 | 12,48 | 53,1 | 16,16 | 124,5 | 0 |

В ней каждый столбец соответствует обеляющему фильтру, настроенному на сигнал $X_r$, а каждая строка – входному сигналу $X_v$. Здесь нулевые значения ВИР – это признак эквивалентности сигналов в теоретико-информационном смысле. Проблемные пары ЭРЕ выделены в таблице полужирным шрифтом. В частности, это две фрикативные фонемы |щ| – |ч|.

Видно, что даже в худшем случае значение ВИР находится в пределах 0,12...0,20. Причем, полученный результат практически не зависит от диктора и реализации ЭРЕ. А при увеличении порядка $p$ до 20…30 минимальное значение ВИР возрастает до 0,20…0,30.

Описанный эксперимент был проведён сто раз по отношению к $J_r = 100$ независимым реализациям ЭРЕ из указанной проблемной пары. При этом объём выборки $n_r$, $r = \overline{1,R}$ по каждой их реализации был сокращен до приемлемых в задачах АРР 120 отсчётов. Сначала из множества реализаций фонемы |ч| согласно правилу (11) была отобрана её эталонная реализация. Затем по ней методом Берга-Левинсона была построена АР-модель (12) порядка $p = 20$ и соответствующий обеляющий фильтр (14). Выборочная дисперсия его отклика с точностью до константы

и множителя 0,5 определяет, согласно выражению (15), значение решающей статистики по МОФ (16). Поэтому подавая на вход ОФ сто разных реализаций фонемы |ч|, мы получали последовательность реализаций минимальной решающей статистики, или статистики МИР. Её график показан на рисунке 1, кривая 1. Здесь по оси абсцисс отложены номера реализаций. Для сравнения на том же рисунке представлена кривая 2 – это последовательность разных реализаций альтернативного варианта решающей статистики (16) при подаче на вход ОФ множества реализаций фонемы |щ|.

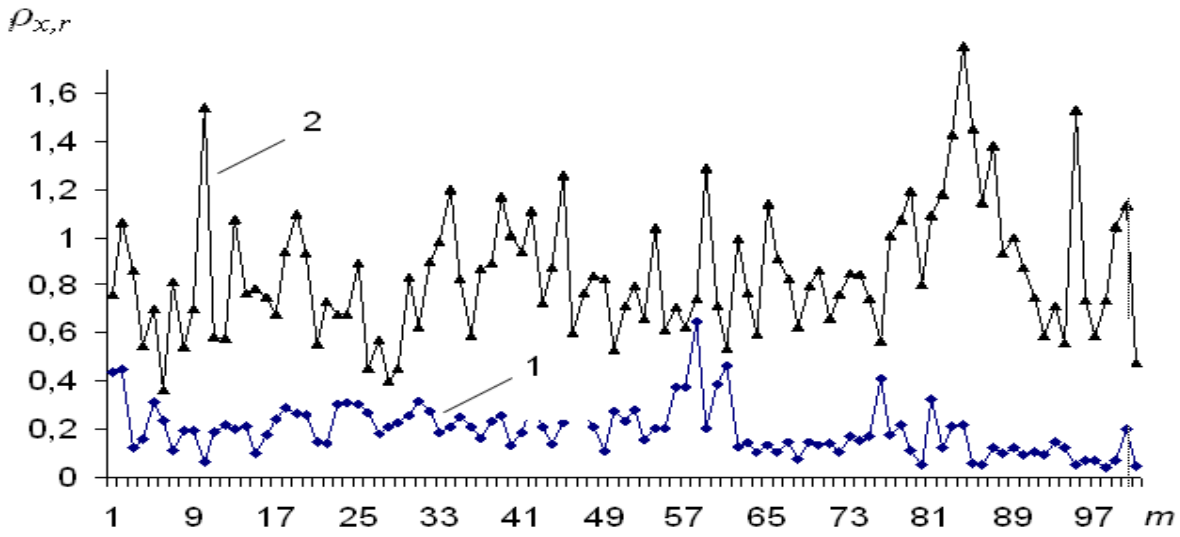



Рисунок 1. Решающая статистика МИР

Видно, что между двумя кривыми в среднем существует значительный «зазор», что говорит о надёжном, даже в проблемном случае, распознавании по МОФ двух рассматриваемых ЭРЕ. При этом полученные результаты хорошо сочетаются, даже визуально, с моделью стационарного временного ряда.

Для полноты картины на рисунке 2 представлена гистограмма распределения решающей статистики МИР (16) из рассматриваемого примера, рассчитанная по кривой 1 на рисунке 1.

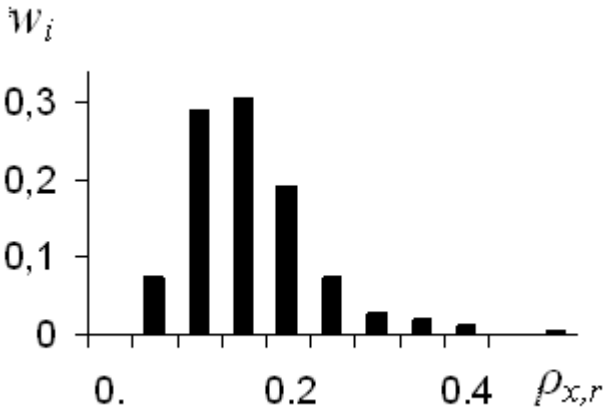



Рисунок 2. Гистограмма распределения решающей статистики МИР

Нетрудно заметить, что качественно она имеет вполне регулярный вид, напоминающий $\chi^2_L$-распределение с некоторым числом $L$>>1 степеней свободы.

Для подтверждения и развития этой идеи на рисунке 3 построена (выделена чёрным) гистограмма распределения $L$-кратной дисперсии отклика ОФ (14) $L\sigma_r^2(\mathbf{x})$ на множество реализаций фонемы |щ|. Здесь же для сравнения (серый цвет) представлена теоретическая гистограмма распределения $\chi^2_L$ при равенстве $L = 106 \approx n_r = 120$. Даже невооруженным глазом видно, что обе гистограммы почти идеально согласуются друг с другом.

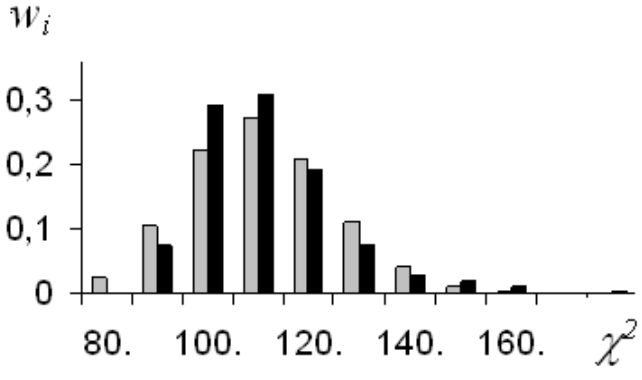



Рисунок 3. Гистограмма распределения статистики $L\sigma_r^2(\mathbf{x})$

Причем, законы распределения статистики МИР для всех других фонем русского языка мало отличаются от рассмотренного в примере. Это видно из рисунка 4, на котором представлены гистограммы распределения

минимальной решающей статистики (16) для гласных |a|, |o| и |y|, полученные при неизменных значениях параметров автоматической обработки сигналов.

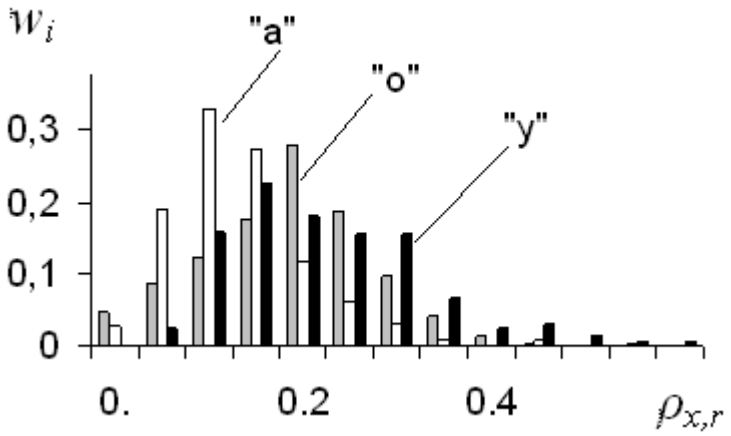



Рисунок 4. Гистограммы распределения решающей статистики МИР для гласных фонем

Таким образом, выводы проведённых экспериментальных исследований в отношении высокой точности $\chi^{2}$-аппроксимация при обработке речи являются веским обоснованием как гауссовой модели речевых сигналов, так и собственно МОФ в разнообразных задачах АОР.

### 2.4. Обсуждение теоретической части

Таким образом, благодаря ИТВР и МОФ удалось в достаточной степени четко сформулировать основную идею критерия минимума информационного рассогласования на основе метода обеляющего фильтра и подтвердить её на двух наиболее актуальных задачах автоматической обработки речи. В перспективе данный подход может найти применение как инструмент предварительной обработки и группирования множества распознаваемых речевых сигналов по принципу их теоретически достаточной различимости. Например, в задачах голосового управления с ограниченным набором команд и используемых слов всегда существует возможность замены каждого отдельного слова на его синоним, если это

сопровождается увеличением потенциальной способности системы АОР к его распознаванию в смысле максимизации минимальной величины ИР.

Если говорить об основных ограничениях на оптимальность и область применения предложенного критерия, то, прежде всего, они связаны с предположением о гауссовском распределении речевых сигналов в постановочной части рассмотренных задач. Однако указанное ограничение только на первый взгляд кажется чрезмерно жестким. В действительности закон $\mathbf{P}_X$ может существенно отличаться от гауссовского. В таком случае применяемая в данной работе модель $N(\mathbf{K}_X)$ определяет его ортопроекцию на семейство гауссовских распределений как наиболее близкое приближение в теоретико-информационном смысле. В работе [4] к тому же показано, что гауссовский закон в его адаптивном варианте теоретически строго обусловлен общесистемным принципом максимальной энтропии наблюдений для широкого круга задач и алгоритмов, использующих моменты распределения не выше второго порядка. Отметим, что подавляющее большинство известных алгоритмов АОР относится именно к этому кругу, т. к. основывается, главным образом, на методах корреляционной и спектральной обработки сигналов.

Отдельного внимания заслуживает вопрос о методах адаптации обеляющего фильтра. В настоящей работе рассмотрен наиболее простой из них, сводящийся к непосредственному оцениванию АКМ по формуле (2). Во многих случаях такой подход вполне обоснован достаточно большими объёмами повторных выборок наблюдений $\{\overrightarrow{x\ e}\}$. В условиях же малых выборок наблюдений, например, в задачах пофонемной обработки речи, предпочтение следует отдавать более скоростным методам авторегрессионого анализа, таким как методы Берга, Юла-Уолкера и другие. Отметим, что все они используют обеляющий фильтр в роли ключевого элемента осуществляемой обработки информации.

# 3. Информационная система фонетического анализа речи

В данной главе подробно описана структура и принцип действия принадлежащей авторам данного учебного пособия и зарегистрированной от их имени в Роспатенте информационной системы (ИС) фонетического анализа речи (ФАР). Особенностью данной ИС является использование критерия МИР в метрике Кульбака-Лейблера, представленного в частотной области. Главное назначение ИС – автоматическое выделение из непрерывного потока речи элементарных (минимальных) речевых единиц (ЭРЕ) типа фонем и их аллофонов. В определенном смысле она представляет собой фонетический анализатор речи.

Примеры практического применения фонетического анализатора можно привести из самых различных областей. Это может быть, например, задача анализа качества речи по её фонетическому составу, как для отдельного диктора, так и для группы дикторов, народности или даже всего национального языка. В качестве прикладной задачи можно привести для примера сжатие речевых данных и их высокоскоростную передачу по каналу связи при сохранении высокого качества речи.

Главной проблемой при построении такого рода анализаторов до настоящего времени было отсутствие чёткого определения самого понятия фонемы. Существует множество её определений, однако зачастую они расплывчаты и неточны, а подчас даже противоречат друг другу. Особенностью данной ИС является применение теоретико-информационного подхода и метода обеляющего фильтра в расчете на АР-модель ЭРЕ с нормировкой по дисперсии порождающего шума. Именно такой подход позволил впервые дать строгое с математической точки зрения определение фонемы – *фонема представляет собой кластер, составленный из множества реализаций однотипных ЭРЕ при минимальных различиях между ними в метрике Кульбака-Лейблера.*

И именно это определение было положено авторами в основу рассматриваемой далее ИС. При этом ИС обеспечивает выполнение следующих функций:

- автоматическое выделение фонем из входных данных;
- автоматическая обработка фонем;
- хранение фонем;
- визуализация полученных результатов.

### 3.1. Основные элементы ИТВР

Несмотря на существующие различия в одноименных речевых метках все они воспринимаются человеком как нечто общее, иначе речь утратила бы свою информативность. Можно поэтому утверждать, что одноименные метки-реализации $\mathbf{x}_{r,\,j}$, $j=\overline{1,J_r}$, $J_r>>1$ в сознании человека группируются в соответствующие классы или речевые образы $X_r=\{\mathbf{x}_{r,\,j}\}$, $r=\overline{1,R}$ [2]. Каждый такой класс имеет четко очерченные границы вокруг некоторого центра – эталонной метки данного образа. В информационной теории восприятия речи указанные эталоны определяются в строгом, теоретико-информационном смысле: речевая метка $\mathbf{x}_r^*\subset X_r$ образует информационный центр-эталон $r$-го речевого образа, если в пределах множества $X_r$ она характеризуется минимальной суммой информационных рассогласований по Кульбаку-Лейблеру относительно всех других его меток-реализаций $\mathbf{x}_{r,\,j}$, $j=\overline{1,J_r}$. В какой-то степени это аналог понятия «центр массы» физического тела.

Благодаря понятию информационного центра речевого образа становится очевидным механизм восприятия речи. Анализируемый (слышимый) речевой сигнал $\mathbf{x}$ в пределах конечного множества речевых образов $\{X_r\}$ отождествляется с тем $X_\nu$ из них, которому отвечает минимум информационного рассогласования между вектором $\mathbf{x}$ и

соответствующим эталоном $\mathbf{x}_{\nu}^{*}$, $\nu \le R$. Это известная формулировка критерия минимума информационного рассогласования (МИР) в задаче автоматического распознавания речи (АРР).

Используемый в ИТВР статистический подход в задаче АРР, по-видимому, не имеет альтернатив ввиду отмеченной выше проблемы вариативности речи. Однако сама идея статистического (по ансамблю реализаций) усреднения сигналов наталкивается здесь на ряд принципиальных препятствий. Прежде всего, это особенности речеобразующего механизма человека. У разных людей он сильно разнится по параметрам, например, по частотным характеристикам своей модели «акустической трубы». Как результат, одноимённые речевые метки, даже элементарные речевые единицы типа отдельных фонем, имеют существенно разные автокорреляционные свойства. Какое-либо усреднение таких сигналов между собой в принципе не обоснованно. Иными словами, понятие речевого образа оказывается не эквивалентным понятию статистического класса выборок с однородными распределениями.

### 3.2. Архитектура информационной системы

На рисунке 5 изображена блок-схема разработанной информационной системы фонетического анализа слитной речи.

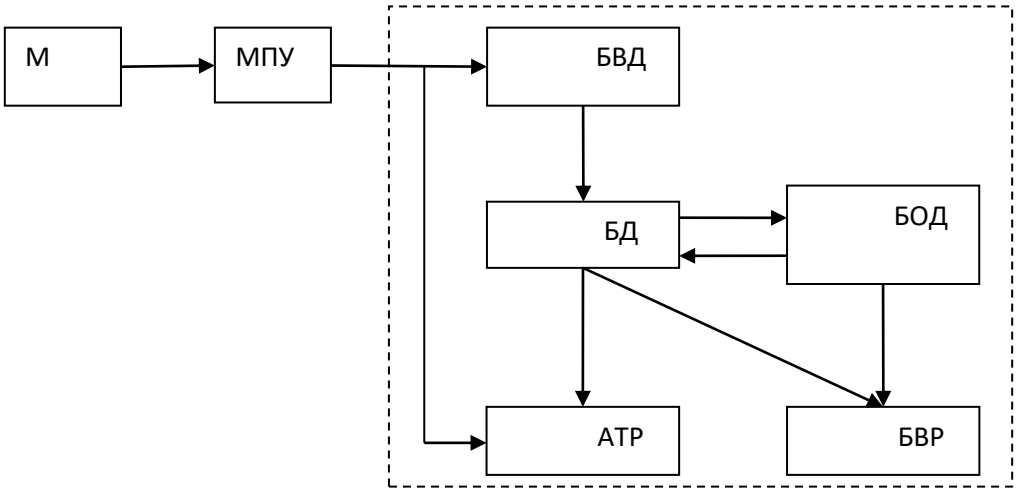



Рисунок 5. Блок-схема разработанной информационной системы фонетического анализа слитной речи

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 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 поступать в БВР через БОД.

*Структура блока ввода данных*

Данный блок предназначен для считывания звуковых данных из файлов, их обработки и записи в БД. На рисунке 6 изображена его структурная схема.

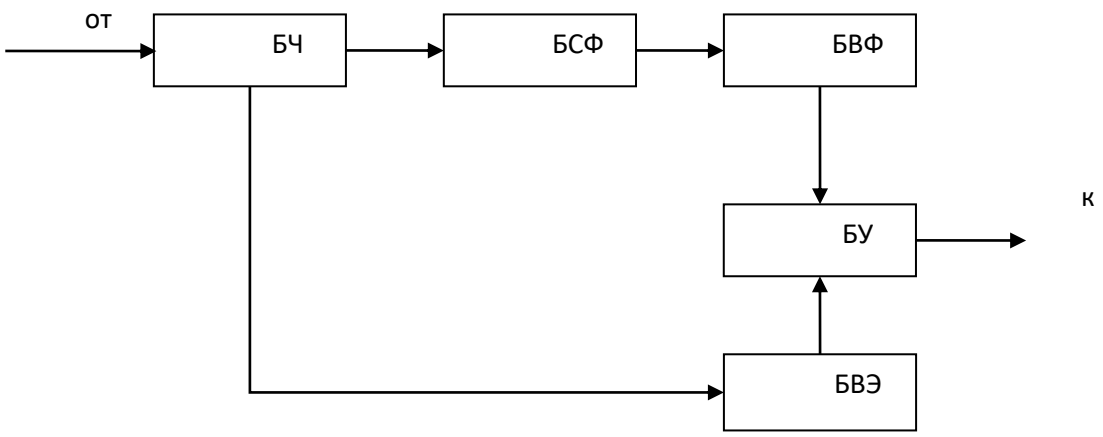



Рисунок 6. Структурная схема блока ввода данных

Здесь:

– БЧ – блок чтения;

– БСФ – блок сегментирования на фонемы;

– БВФ – блок выбора фонем;

– БУ – блок управления;

– БВЭ – блок выбора эталонов.

Ввод данных может осуществляться в следующих режимах:

1) непосредственного ввода данных;
2) пакетной обработки данных;
3) обработки фонем.

Блок чтения производит считывание информации из одного (в режиме 1) или нескольких (в режимах 2 и 3) звуковых файлов и передает её в зависимости от режима работы либо в БСФ, либо в БВЭ.

В том случае, если ввод данных осуществляется в режимах 1 или 2, информация от БЧ поступает в БСФ, который производит сегментирование входного сигнала на фонемы и вычисляет их векторы АР-коэффициентов при помощи рекуррентной процедуры Берга-Левинсона. После этого полученный набор векторов передаётся в БВФ, в котором производится их кластеризация и нахождение информационных центров кластеров по методу переменного дерева. АР-коэффициенты, соответствующие

информационным центрам поступают в БУ, который производит их запись в БД. При работе в режиме 3 каждый звуковой файл рассматривается как реализация соответствующей фонемы. Данные, прочитанные из файла, поступают в БВЭ, который производит выбор эталонной реализации каждой фонемы и вычисление её АР-коэффициентов. После этого полученные коэффициенты передаются в БУ, который производит их запись в БД.

*Структура блока обработки данных и вывода результатов*

Данный блок предназначен для организации работы с группами дикторов, экспорта данных из БД и визуализации спектров фонем. На рисунке 7 изображена его структурная схема.

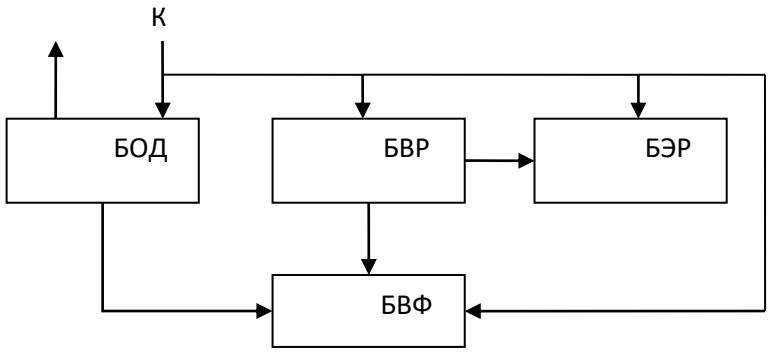



Рисунок 7. Структурная схема блока обработки данных и вывода результатов

Здесь БОД – блок обработки данных. Данный блок предназначен для организации работы с группами дикторов: осуществляет формирование группы либо по заданным пользователем критериям (пол, возраст, регион проживания), либо из выбранных пользователем дикторов; определяет информационные центры для каждой группы фонем. БВР – блок вывода результатов. Данный блок позволяет выводить на экран спектры фонем, а также их АР-коэффициенты. БЭР – блок экспорта результатов позволяет экспортировать наборы АР-коэффициентов отдельных дикторов или групп в текстовый файл. БВФ – блок воспроизведения фонем позволяет выводить

синтезированный звуковой сигнал, соответствующий выбранной фонеме на звуковую карту компьютера. Пользователь может самостоятельно изменять параметры воспроизведения.

Подсистема *блока автоматического транскрибирования речи* предназначена для автоматического выделения фонем из устной речи, сопоставление их с фонемами выбранного диктора и анализа получаемых результатов.

Подсистема позволяет:

- прослушивать звуковые файлы в формате **.wav*;
- просматривать речевой сигнал в различном масштабе времени;
- автоматически выделять фонемы из входных данных;
- производить спектральный анализ полученных фонем;
- экспортировать АР-коэффициенты фонем в текстовый файл.

На рисунке 8 изображена блок-схема подсистемы автоматического выделения фонем.

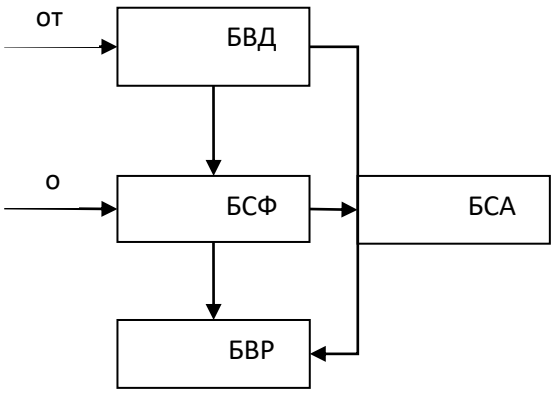



Рисунок 8. Блок-схема подсистемы автоматического выделения фонем

Здесь:

- ИС ФБД – информационная система «Фонетическая база данных русского языка»;
- БВД – блок ввода данных;
- БСФ – блок сегментирования входного сигнала на фонемы;

– БВР – блок вывода результатов;

– БСА – блок спектрального анализа.

Блок ввода данных предназначен для считывания информации из звуковых файлов и передачи её в блоки вывода результатов и сегментации фонем.

Блок сегментирования входного сигнала на фонемы предназначен для автоматического выделения фонем из речевого сигнала и их сопоставления с фонемами выбранного диктора. Выделение фонем производится в соответствии с алгоритмом ИТВР [1; 2].

Блок вывода результатов осуществляет отображение и прослушивание загруженного сигнала, а также выделенных в нём фонем.

Блок спектрального анализа позволяет отображать спектральную плотность мощности (СПМ) выбранной фонемы в различных режимах. Кроме того, возможен вывод полученных результатов на печать.

*Схема базы данных*

Разработанная информационная система использует базу данных формата *Microsoft Access*. На рисунке 9 показана схема этой базы данных.

БД включает 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 набор фонем дикторов.

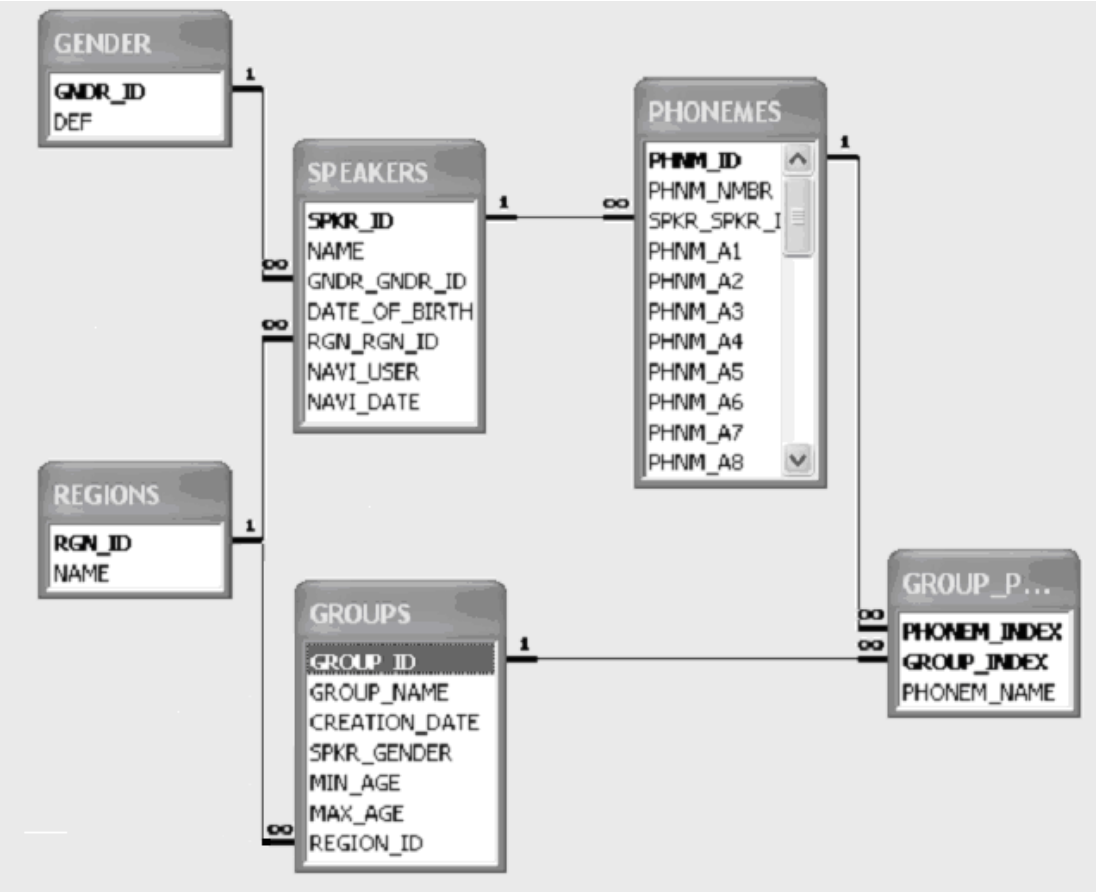



Рисунок 9. Схема базы данных формата *Microsoft Access*

- GROUPS – таблица хранит информацию о группах дикторов: название группы, дату создания, пол дикторов, входящих в группу, максимальный и минимальный возраст дикторов, регион.
- GROUP_PHONEMES – таблица служит для организации связи «многие-ко-многим» между таблицами PHONEMES и GROUPS, содержит наименование фонем данной группы, определяемых пользователем на слух.

### 3.3. Описание программы для ПК

*Назначение и основные возможности программы*

Данная программа предназначена для проведения экспериментальных исследований в области автоматического распознавания речи. Программа реализована в среде разработки MATLAB 6.5.

Программа позволяет:

- Проводить автоматическое сегментирование речевого сигнала на фонемы, находить среди них чёткие ЭРЕ, рассчитывать по ним коэффициенты авторегрессионной модели и записывать полученные результаты в ФБД.
- Отображать содержимое БД.
- Синтезировать вокализованные и невокализованные фонемы по их АР-коэффициентам.
- Формировать группы дикторов по заданному критерию.
- Экспортировать АР-модели фонем из БД в текстовый файл.

Программа имеет удобный и понятный графический интерфейс, реализованный средствами системы MATLAB 6.5. В качестве исходных данных служат стандартные звуковые WAV.

Рекомендуемые требования к аппаратному и программному обеспечению персонального компьютера: Pentium IV, 256 Мбайт ОЗУ, Windows XP, Matlab 6.5. Для запуска приложения необходимо запустить среду разработки *Matlab*, интерфейс которой показан на рисунке 10.

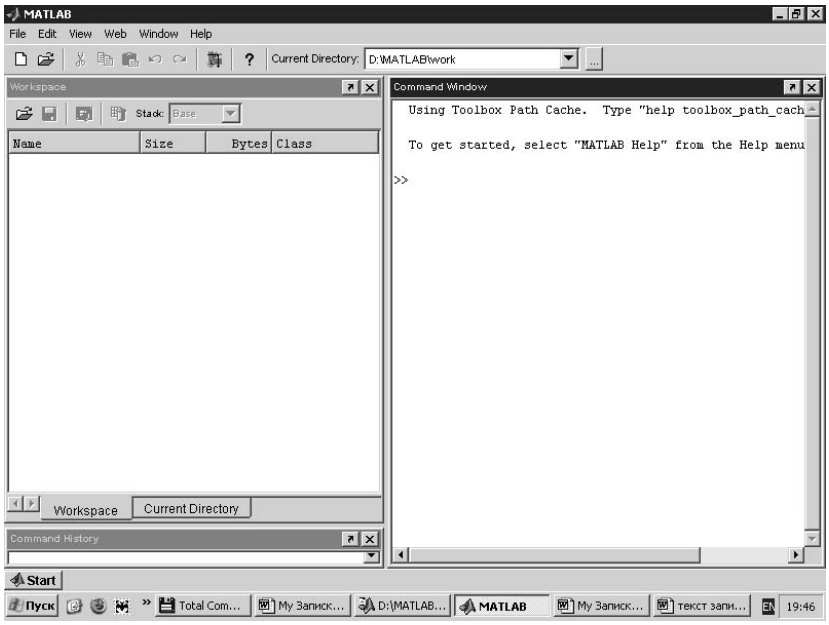



Рисунок 10. Интерфейс среды разработки *Matlab*

Для корректного выполнения в поле *Current Directory* необходимо указать путь к каталогу с программой (например, “D:\MATLAB\work”). Далее необходимо запустить файл “phonemexplorer.m”. Для этого нужно открыть нужный файл в *Matlab* и нажать кнопку F5. Появится главное окно, изображенное на рисунке 11.

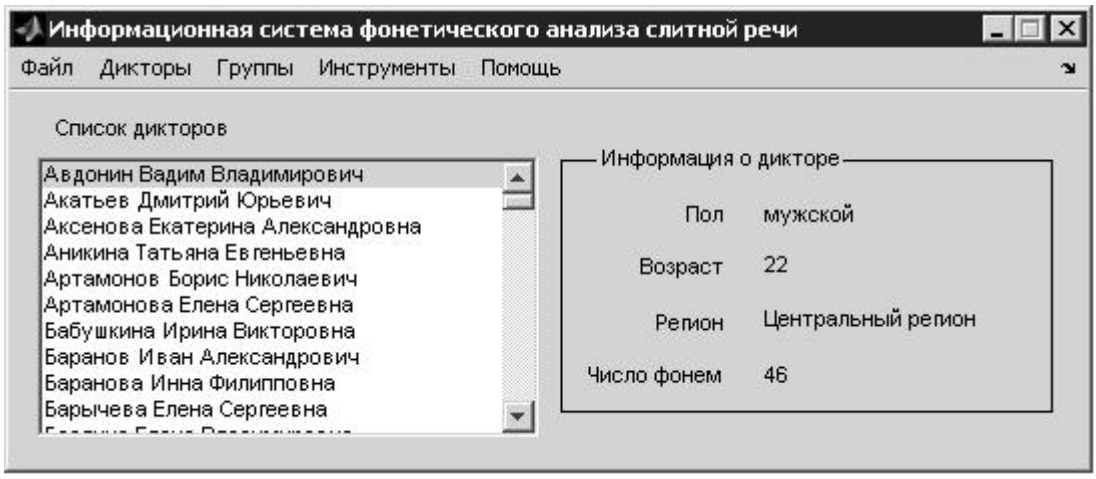



Рисунок 11. Главное окно *Matlab*

На этой форме отображаются дикторы, внесенные в БД и главное меню программы. При выборе любого диктора из списка в правой части окна отображается краткая информация о нём. 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. Подробнее об этом будет рассказано в разделе 3.2 «Режимы ввода данных».

- Меню «Группы» содержит функции для работы с группами дикторов:
  - "Показать" – открывает форму для работы с группами дикторов. Подробнее об этом будет рассказано в разделе 3.3 «Работа с группами».
- 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.

*Режимы ввода данных*

В данном разделе описывается ввод новых дикторов в систему и работа с существующими дикторами, а также ввод фонем.

Для ввода нового диктора выбирается меню «Дикторы»>«Добавить». При этом на экране появляется форма, показанная на рисунке 12.

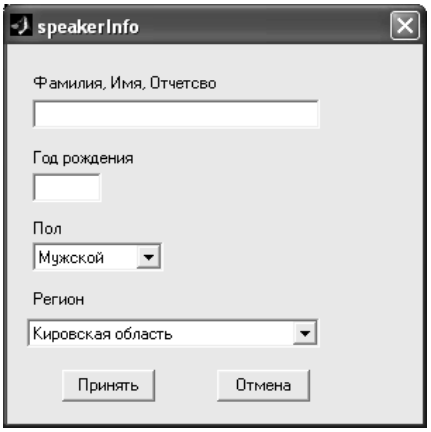



Рисунок 12. Функция добавления нового диктора в БД

Пользователю необходимо ввести фамилию, имя и отчество диктора, год его рождения, выбрать регион, в котором проживает диктор. При нажатии кнопки «Принять» в БД будет добавлен новый диктор, и его имя будет отображаться в списке дикторов. Нажатие кнопки «Отмена» закрывает форму без сохранения введенных данных.

Для того чтобы изменить информацию о каком-либо дикторе, необходимо сначала выбрать его в списке дикторов, а затем выбрать меню «Дикторы»>«Изменить». В этом случае на экране появится форма, аналогичная показанной на рисунке 12. Однако в её полях будут содержаться сведения, соответствующие выбранному диктору. После внесения необходимых изменений и нажатия кнопки «Принять» изменения будут записаны в БД. Нажатие кнопки «Отмена» закрывает форму без сохранения изменений.

Ввод фонем возможен как из отдельных, так и сразу из нескольких звуковых файлов. Для того чтобы добавить фонемы конкретного диктора, необходимо выделить его в списке дикторов и выбрать команду меню «Дикторы»>«Фонемы»>«Добавить». В том случае, если у диктора уже есть какие-либо фонемы, на экране появится предупреждение, показанное на рисунке 13.

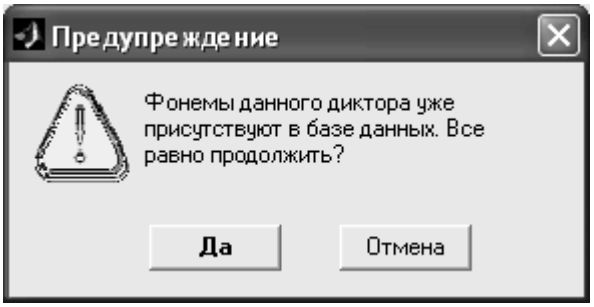



Рисунок 13. Предупреждение о присутствии диктора в БД

При нажатии кнопки отмена «Да» новые фонемы будут добавлены к уже существующим фонемам. При нажатии кнопки «Отмена» ввод фонем будет отменён.

При выполнении операции по вводу фонем на экране появится стандартный системный диалог открытия файлов. При вводе фонем из отдельного звукового файла в списке «Тип файлов» требуется выбрать «Звуковые файлы (*.wav)» (выбран по умолчанию) или «Все файлы (*.*)». В этом случае после нажатия кнопки «Открыть» начнётся процедура анализа выбранного звукового файла и запись АР-моделей выделенных из него фонем в БД.

В случае ввода фонем из нескольких звуковых файлов в списке "Тип файлов" требуется выбрать «Файлы речевых баз (*.txt)» и выбрать специальным образом сформированный текстовый файл, содержащий информацию о расположении звуковых файлов. После выбора файлов на экране появится форма выбора режима ввода, показанная на рисунке 14.

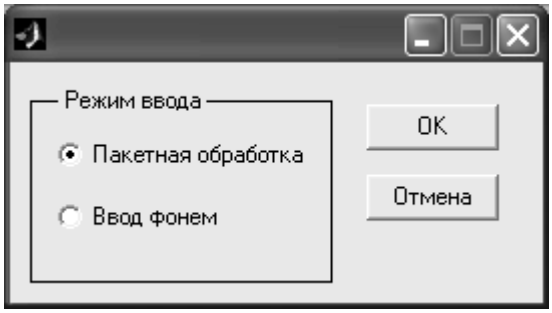



Рисунок 14. Форма выбора режима ввода

При выборе режима «Пакетная обработка» каждый звуковой файл, перечисленный в файле речевой базы, будет прочитан и сегментирован на фонемы. После этого производится разбиение выделенных фонем на классы и запись АР-моделей, соответствующих их информационным центрам в БД.

При выборе режима «Ввод фонем» система предполагает, что речевая база состоит из нескольких папок, в каждой из которых содержится несколько реализаций одной фонемы. Среди этих реализаций выбирается эталонная, и её АР-модель записывается в БД.

Для того, чтобы экспортировать фонемы выбранного диктора в текстовый файл, необходимо выбрать меню «Дикторы»>«Фонемы»>«Экспорт…». В этом случае на экране откроется стандартный системный диалог сохранения файла, в котором нужно выбрать или ввести имя файла, в который будет осуществлен экспорт численных значений коэффициентов авторегрессии. Запись данных осуществляется в текстовом формате. В качестве разделителя используется символ ';'. Такой файл может быть прочитан большинством современных электронных таблиц, например *Microsoft Excel*.

*Работа с группами*

Группа дикторов представляет собой множество фонем, сформированное из фонем нескольких дикторов. Причём эти дикторы отбираются по заданным критериям. На рисунке 15 оказана форма работы с группами.

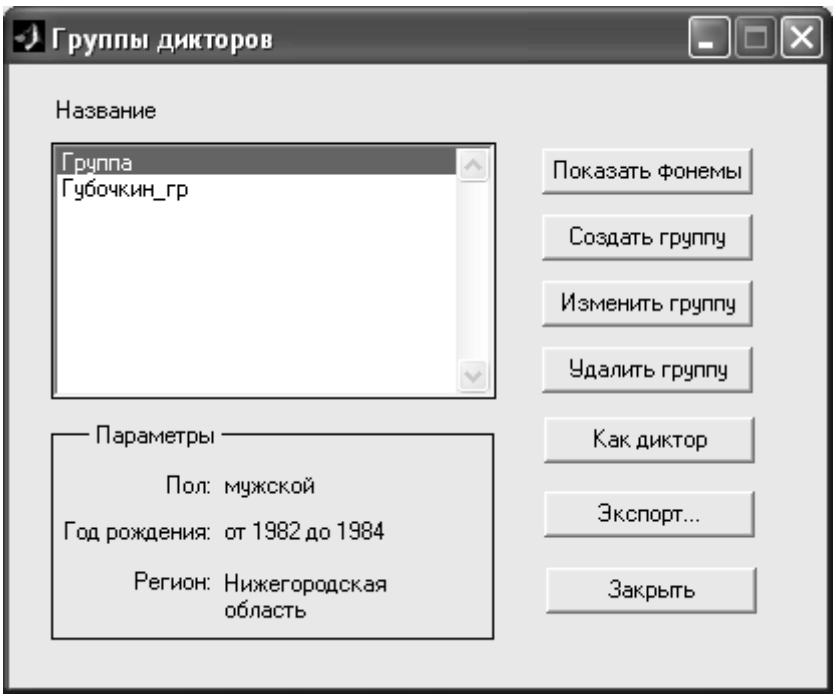



Рисунок 15. Форма работы с группами

Чтобы просмотреть фонемы выбранной группы, необходимо нажать кнопку «Показать фонемы». При этом на экране откроется форма

просмотра фонем. Более подробно о ней будет рассказано в следующем разделе.

Для создания новой группы необходимо нажать кнопку «Создать группу». При этом на экране появится форма, в которой можно выбрать способ отбора дикторов для новой группы. Вид этой формы показан на рисунке 16.

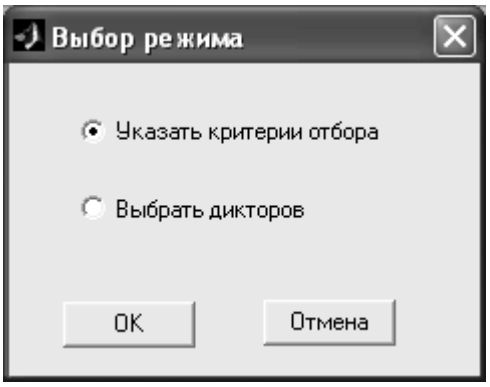



Рисунок 16. Форма способа отбора дикторов для новой группы

В случае выбора режима «Указать критерии отбора» в группу будут включены только те дикторы, информация о которых соответствует заданным критериям. На рисунке 17 показана форма для ввода значений критериев отбора. На этой же форме вводится и название группы.

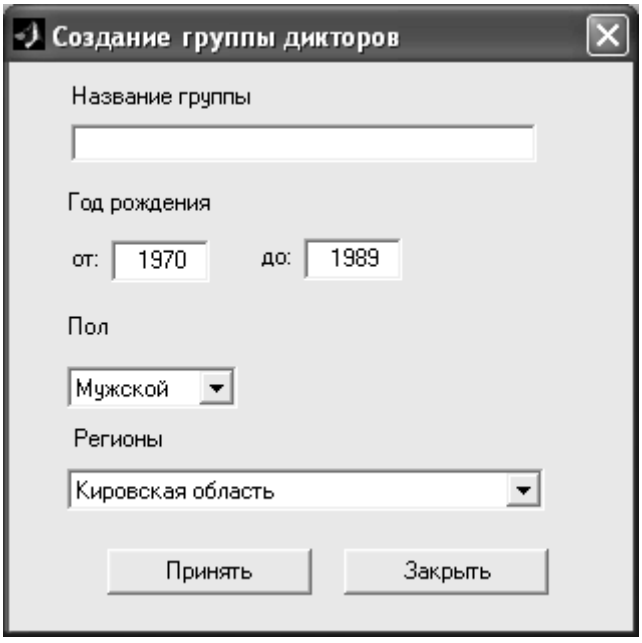



Рисунок 17. Форма для ввода значений критериев отбора

После задания значений всех критериев для создания новой группы необходимо нажать кнопку «Принять». В этом случае производится создание новой группы с соответствующим  именем и наполнение её фонемами.

При выборе режима «Выбрать дикторов» на экране появляется форма выбора дикторов из всех имеющихся в списке. Вид этой формы показан на рисунке 18.

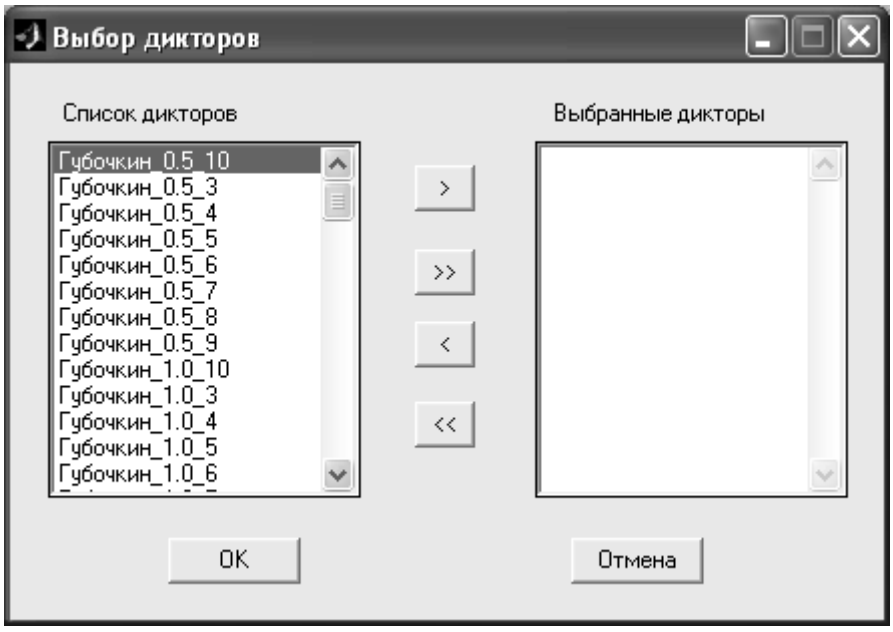



Рисунок 18. Форма выбора дикторов из имеющихся в списке

Нажатие на кнопку переносит выделенного диктора в список выбранных дикторов. Кнопка позволяет выбрать сразу всех дикторов. Кнопка позволяет перенести одного выбранного диктора обратно в список дикторов. Кнопка переносит сразу всех выбранных дикторов обратно в список дикторов.

После того как процедура выбора завершена, для создания новой группы необходимо нажать кнопку «ОК». На экране появляется форма с предложением ввести имя новой группы. Вид этой формы показан на рисунке 19.

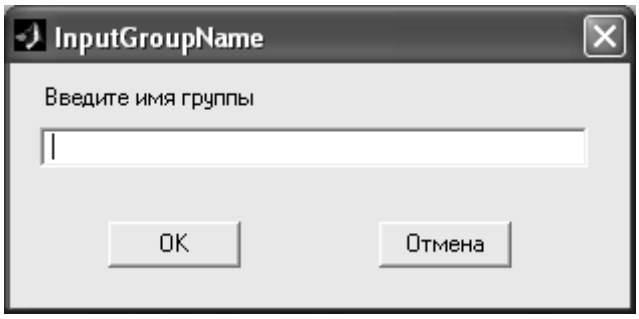



Рисунок 19.

После ввода имени группы производится создание новой группы и наполнение ее фонемами. При добавлении новой группы её название появится в списке групп.

Для изменения критериев отбора дикторов в уже существующей группе необходимо выделить группу в списке групп и нажать кнопку «Изменить группу». В этом случае на экране появится форма, аналогичная показанной на рисунке 17. Однако в её поля будет занесена информация соответствующая выбранной группе. После её изменения и нажатия кнопки «Принять», произведенные изменения будут записаны в базу данных, а наполнение её фонемами будет произведено заново.

Для того чтобы удалить выбранную группу, необходимо нажать кнопку «Удалить группу».

Для того чтобы скопировать группу в таблицу дикторов, необходимо нажать кнопку «Как диктор». В этом случае будет создан новый диктор, имя которого будет соответствовать имени группы, в качестве фонем будут добавлены фонемы, включённые в группу.

Для того чтобы экспортировать коэффициенты авторегрессии, которыми представлены фонемы, включённые в группу в текстовый файл, необходимо нажать кнопку «Экспорт». Экспорт АР-коэффициентов фонем группы аналогичен экспорту АР-коэффициентов фонем отдельного диктора.

Нажатие кнопки «Закрыть» закрывает форму работы с группами.

*Вывод и обработка получаемых результатов*

Разработанная информационная система позволяет отображать спектральную плотность мощности и коэффициенты линейного

предсказания каждой фонемы каждого диктора или группы дикторов. Так же возможен синтез звукового сигнала на основе коэффициентов КЛП с заданной громкостью и частотой основного тона (для вокализованных фонем). На рисунке 20 показана форма синтеза и отображения характеристик фонем.

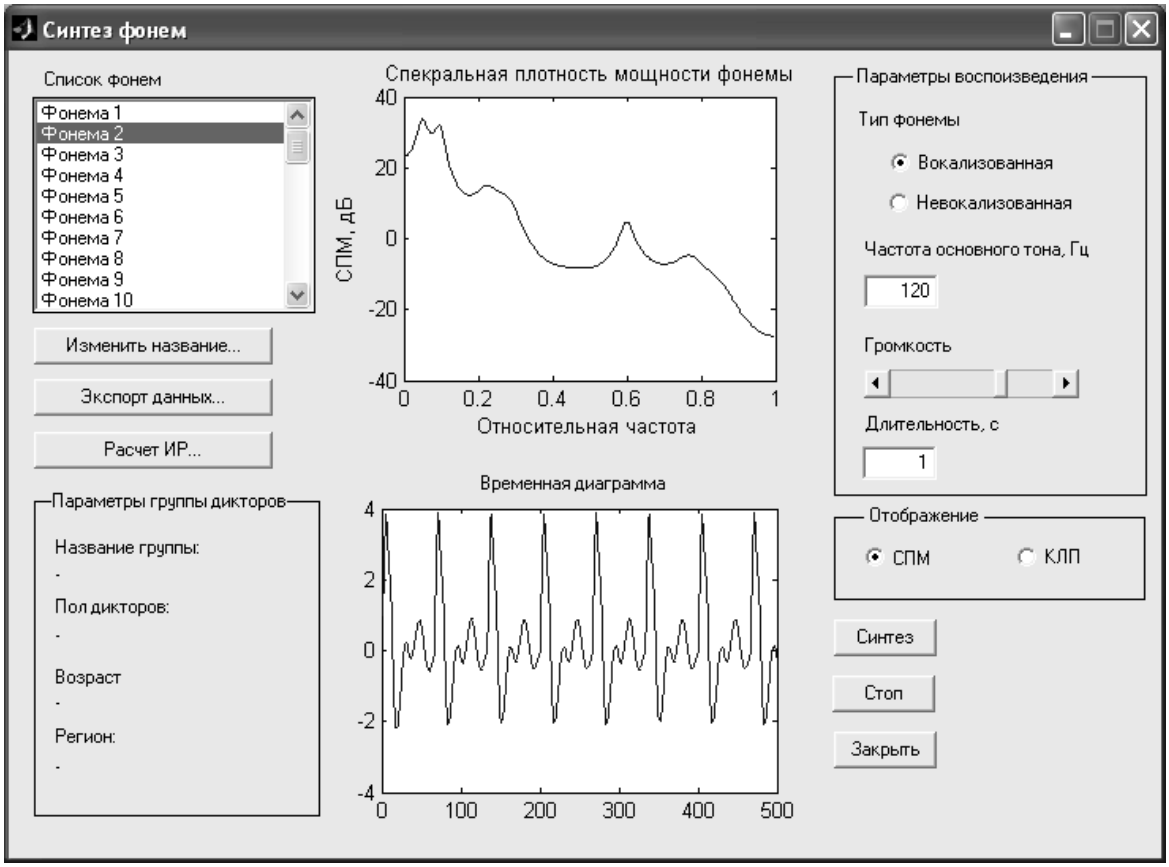



Рисунок 20. Форма синтеза и отображения характеристики фонем

При первоначальном вводе фонем у них отсутствуют названия, и система автоматически присваивает им порядковые номера. Чтобы поименовать выбранную фонему, необходимо нажать кнопку «Изменить название». При этом на экране появится форма, показанная на рисунке 21.

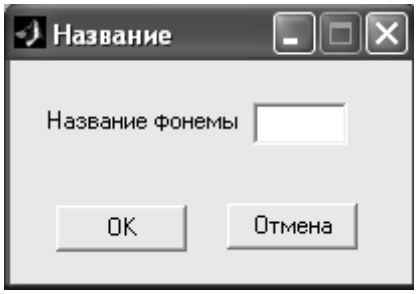



Рисунок 21.

После ввода названия фонемы в соответствующем поле и нажатия кнопки «ОК» оно будет записано в базу данных, а вид списка фонем изменится, как это показано на рисунке 22.

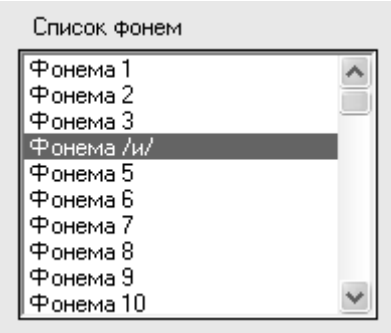



Рисунок 22. Вид списка фонем

Аналогичным образом можно поименовать и другие фонемы. Следует отметить, что фонемы группы дикторов именуются отдельно.

Существует возможность расчета значений матрицы ВИР между фонемами разных дикторов или групп, а также между эталонной фонемой и другими её реализациями. Для этого необходимо нажать кнопку «Расчет ИР». На экране появится соответствующая форма, которая показана на рисунке 23.

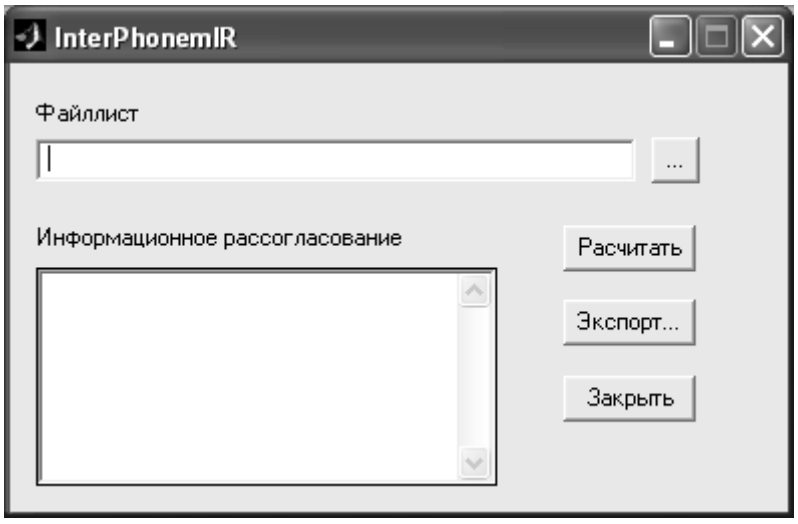



Рисунок 23. Форма расчета ИР

Сначала надо выбрать файл речевой базы данных, введя его полное имя в поле «Файллист» или щелкнув по кнопке ... и выбрав требуемый

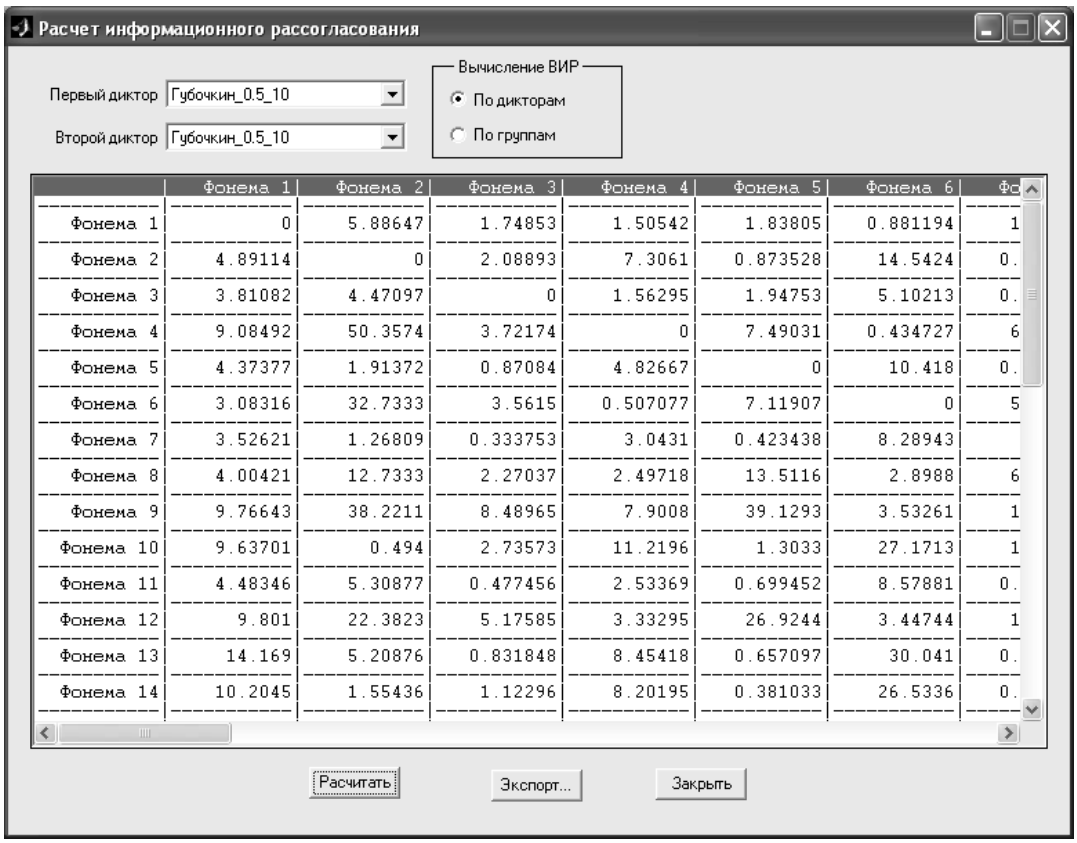
файл с помощью стандартного системного диалога. После этого для расчета ИР необходимо нажать кнопку «Рассчитать». В окне «Информационное рассогласование» появится список значений ИР. Для экспорта его в файл формата *csv* необходимо нажать кнопку «Экспорт». Нажатие на кнопку «Закрыть» закрывает форму.

На рисунке 24 показана форма расчета ВИР между фонемами разных дикторов или групп. Соответствующий режим выбирается радиокнопками «По дикторам» и «По группам».

Для расчета значений матрицы необходимо нажать кнопку «Рассчитать». В окно будет выведена матрица ВИР, строки которой соответствуют обеляющим фильтрам, а столбцы – входным сигналам.

Для экспорта матрицы ВИР в файл формата *csv* необходимо нажать кнопку «Экспорт» и ввести имя файла, в который будут сохраняться данные.

Нажатие кнопки "Закрыть" закрывает форму.




| | Фонема 1 | Фонема 2 | Фонема 3 | Фонема 4 | Фонема 5 | Фонема 6 | Фо |
|---|---|---|---|---|---|---|---|
| Фонема 1 | 0 | 5.88647 | 1.74853 | 1.50542 | 1.83805 | 0.881194 | 1 |
| Фонема 2 | 4.89114 | 0 | 2.08893 | 7.3061 | 0.873528 | 14.5424 | 0. |
| Фонема 3 | 3.81082 | 4.47097 | 0 | 1.56295 | 1.94753 | 5.10213 | 0. |
| Фонема 4 | 9.08492 | 50.3574 | 3.72174 | 0 | 7.49031 | 0.434727 | 6 |
| Фонема 5 | 4.37377 | 1.91372 | 0.87084 | 4.82667 | 0 | 10.418 | 0. |
| Фонема 6 | 3.08316 | 32.7333 | 3.5615 | 0.507077 | 7.11907 | 0 | 5 |
| Фонема 7 | 3.52621 | 1.26809 | 0.333753 | 3.0431 | 0.423438 | 8.28943 | |
| Фонема 8 | 4.00421 | 12.7333 | 2.27037 | 2.49718 | 13.5116 | 2.8988 | 6 |
| Фонема 9 | 9.76643 | 38.2211 | 8.48965 | 7.9008 | 39.1293 | 3.53261 | 1 |
| Фонема 10 | 9.63701 | 0.494 | 2.73573 | 11.2196 | 1.3033 | 27.1713 | 1 |
| Фонема 11 | 4.48346 | 5.30877 | 0.477456 | 2.53369 | 0.699452 | 8.57881 | 0. |
| Фонема 12 | 9.801 | 22.3823 | 5.17585 | 3.33295 | 26.9244 | 3.44744 | 1 |
| Фонема 13 | 14.169 | 5.20876 | 0.831848 | 8.45418 | 0.657097 | 30.041 | 0. |
| Фонема 14 | 10.2045 | 1.55436 | 1.12296 | 8.20195 | 0.381033 | 26.5336 | 0. |

Рисунок 24. Форма расчета ВИР между фонемами разных дикторов или групп

На рисунке 25 показана форма настроек информационной системы. Здесь задаются параметры для работы реализованных в ней алгоритмов.

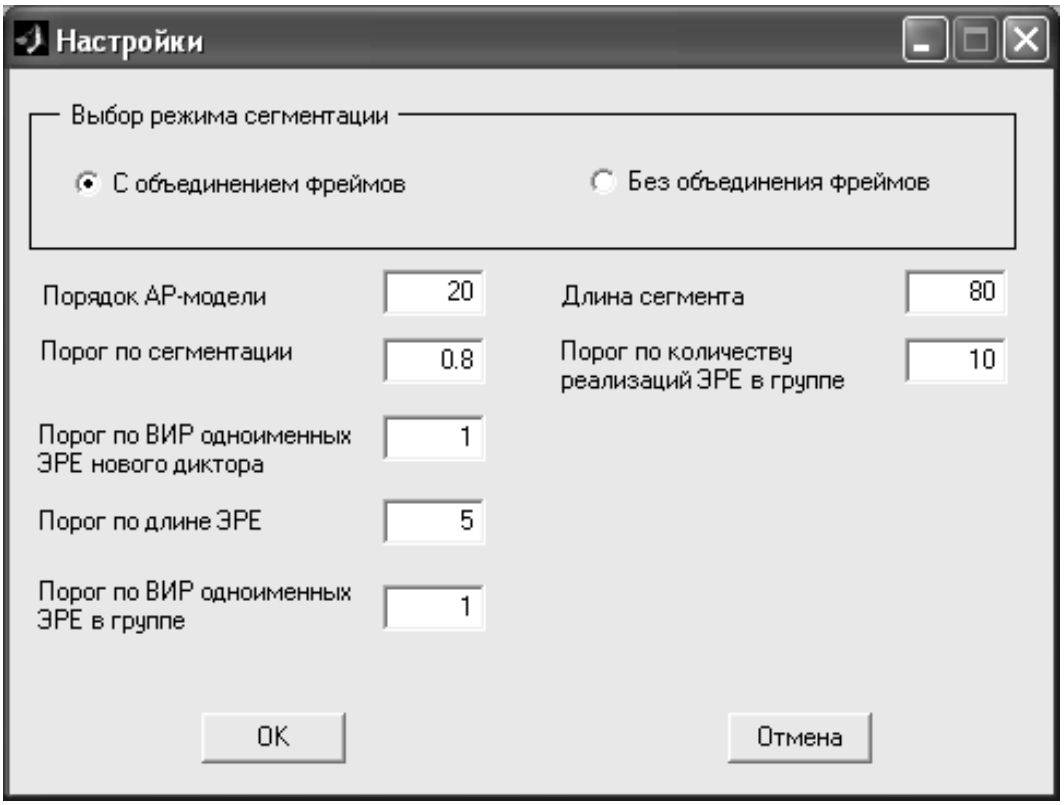



Рисунок 25. Форма настроек информационной системы

Порядок АР-модели – целое число больше 1. 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 ВИР 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 фонема, для того чтобы она могла участвовать в процедуре классификации.

Порог по ВИР 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 отсчётах 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.

*Автоматическое транскрибирование речевого сигнала*

Вход в подсистему автоматического выделения фонем осуществляется через меню «Инструменты»>«Сегментирование» главной формы ИС ФБД (см. рис. 26). В этом случае на экране появятся три формы, показанные на рисунке 27.

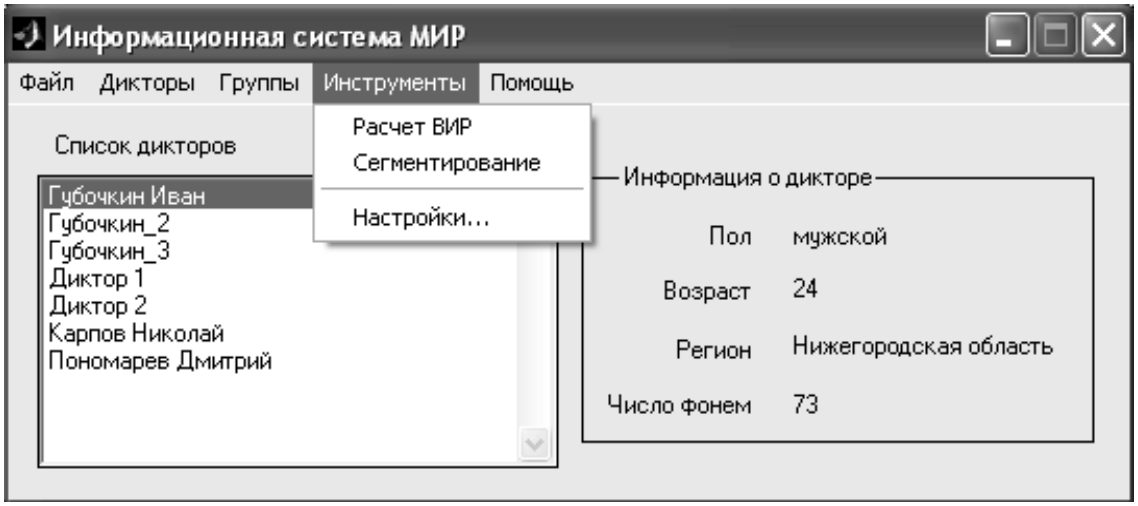



Рисунок 26. Главная форма ИС ФБД

При этом в качестве эталонных будут использоваться фонемы диктора, выбранного из списка на главной форме ИС ФБД.

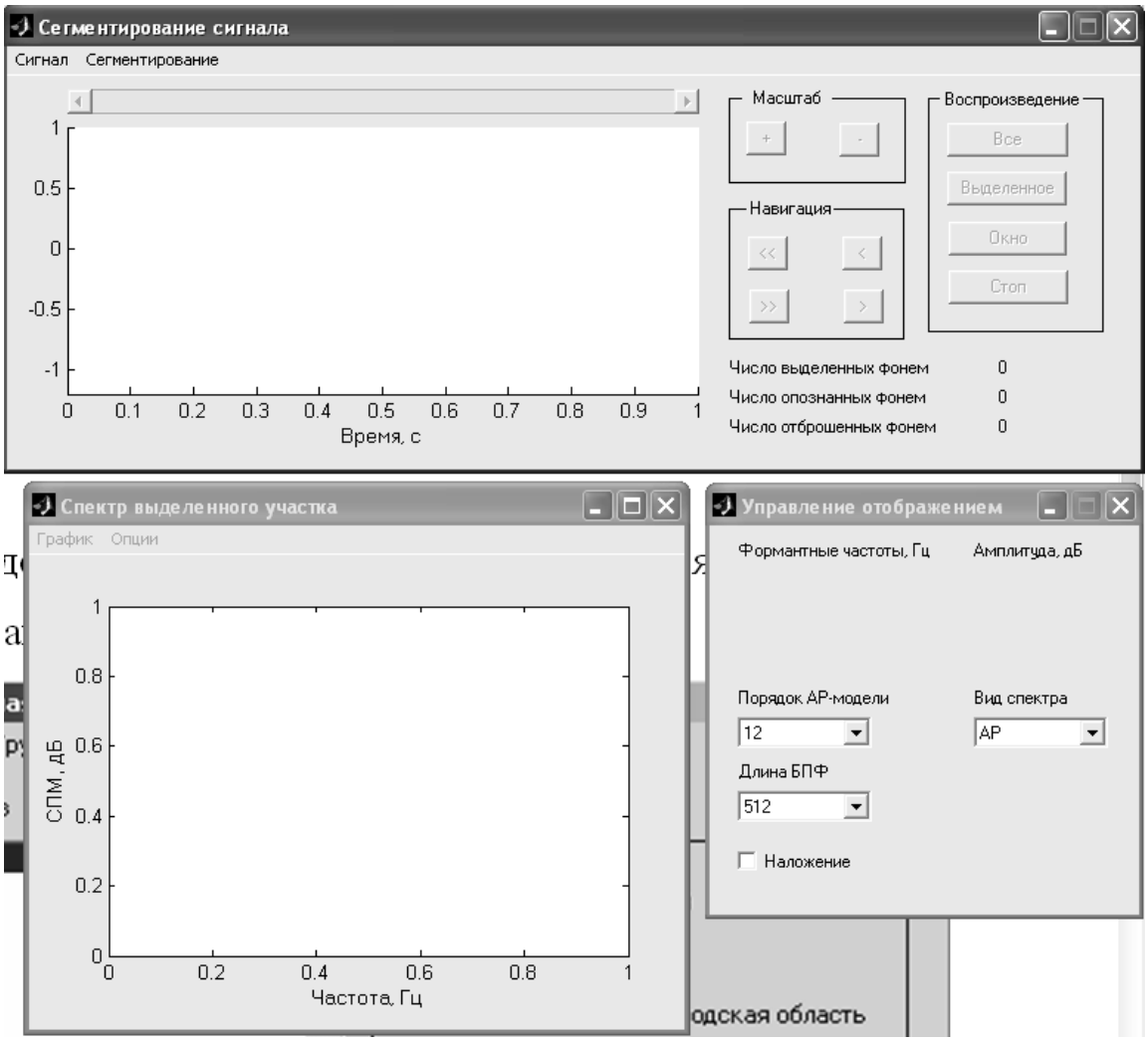



Рисунок 27. Подсистема автоматического выделения фонем

Форма «Сегментирование сигнала» позволяет отображать загруженный сигнал, производить его прослушивание и автоматическое выделение фонем. Ниже приводится краткое описание каждого из разделов главного меню.

- Меню «Сигнал» содержит следующие подменю:
  - «Загрузить» позволяет загрузить звуковой файл в систему для его последующей обработки.
  - «Закрыть» служит для выхода из подсистемы.
- «Сегментирование» содержит меню для проведения процедуры автоматического выделения фонем:
  - «Старт» запускает процедуру выделения фонем.

– «Очистить» удаляет с графика все границы и текстовые обозначения выделенных фонем.

– «Настройки» открывает меню настроек подсистемы.

*Загрузка исходных данных и сегментирование*

Для того чтобы загрузить в программу речевой сигнал, необходимо выбрать меню «Сигнал»>«Загрузить». После этого загруженный сигнал будет отображен на форме «Сегментирование сигнала», которая показана на рисунке 28.

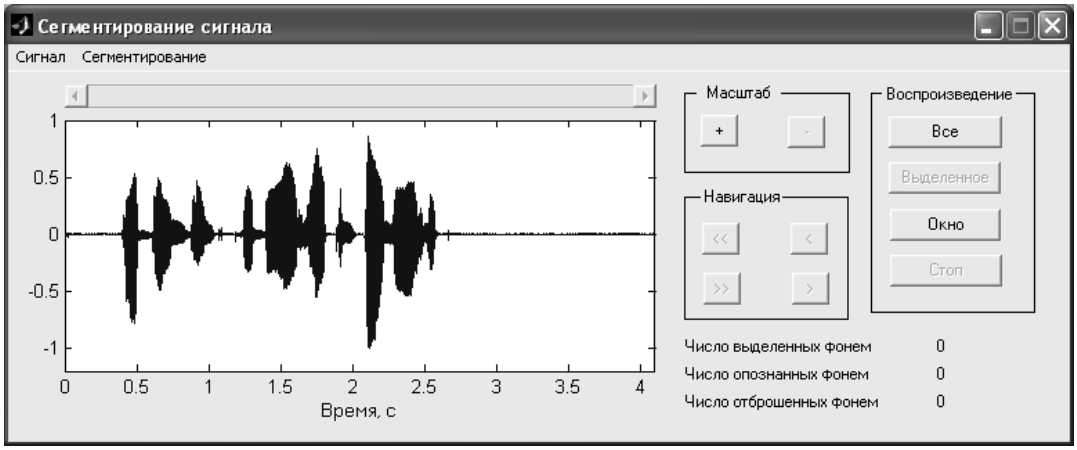



Рисунок 28. Форма «Сегментирование сигнала»

Для изменения масштаба оси времени служат кнопки + и -. Кнопка + позволяет увеличить масштаб в 2 раза, а - – уменьшить. На рисунке 29 показан фрагмент сигнала в увеличенном масштабе.

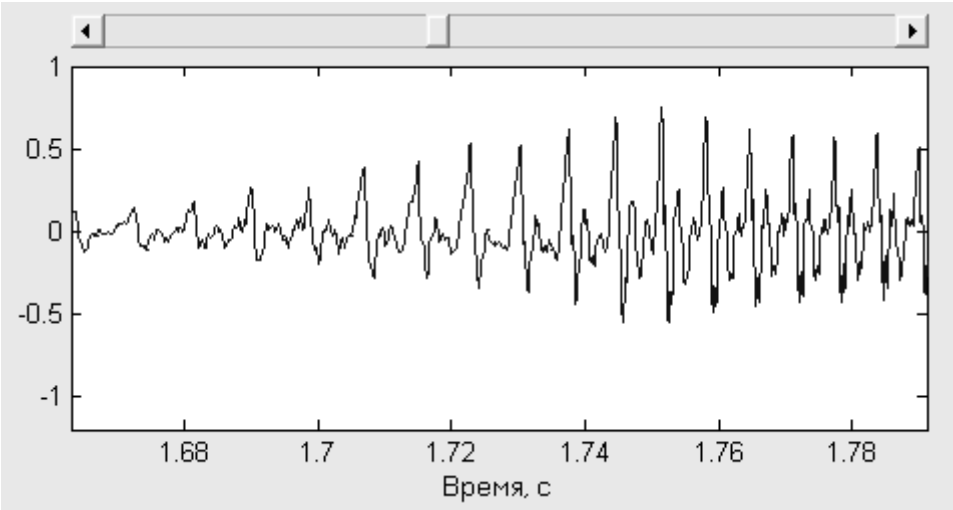



Рисунок 29. Фрагмент сигнала

Навигация по временной диаграмме может осуществляться либо с помощью полосы прокрутки, либо использованием соответствующих кнопок, расположенных в разделе «Навигация». Кнопки << и >> перемещают в начало и конец сигнала соответственно, а кнопки < и > – на один экран влево и вправо.

Программа может воспроизводить речевой сигнал в различных режимах. Кнопка Все воспроизводит загруженный сигнал от начала до конца. Кнопка Выделенное воспроизводит выделенный интервал. Она становится доступной только если выделен какой-либо фрагмент сигнала. Нажатие на кнопку Окно позволяет воспроизвести фрагмент, отображаемый в окне навигации. Нажатие на кнопку Стоп останавливает воспроизведение.

Для того чтобы произвести автоматическое выделение фонем в загруженном сигнале, необходимо выбрать меню «Сегментирование»>«Старт», после чего будет запущена процедура сегментирования. Результат её выполнения показан на рисунке 30.

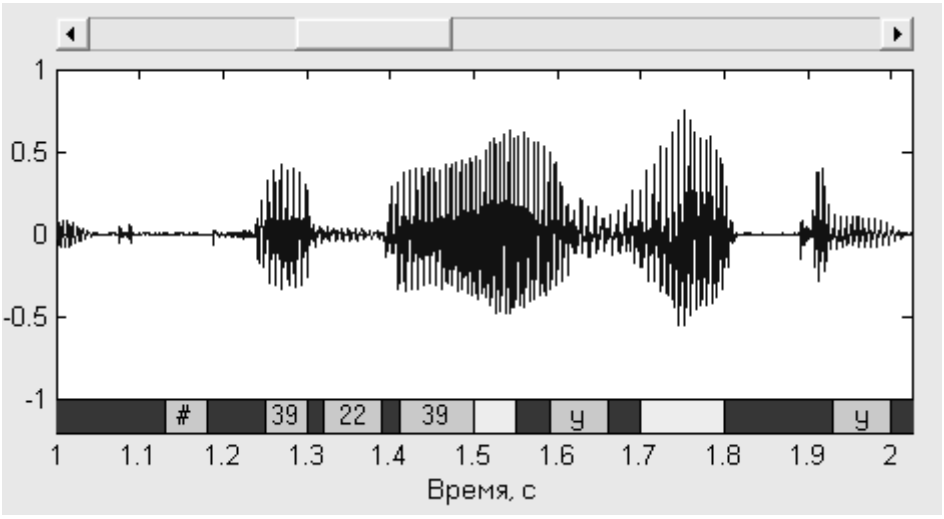



Рисунок 30. Результат выполнения процедуры сегментирования

Здесь красным цветом выделены неинформативные участки. Жёлтым помечены участки, для которых не было найдено подходящей

фонемы выбранного диктора. Сине-зелёным цветом выделены опознанные фонемы. В том случае, если выбранной фонеме присвоено какое-либо имя, оно отображается на диаграмме. Если же никакого имени не присвоено, то отображается только порядковый номер фонемы.

Для того чтобы удалить результаты автоматического выделения фонем, необходимо выбрать меню «Сегментирование»>«Очистить». В этом случае экран приобретет вид, аналогичный показанному на рисунке 28.

При выборе меню «Сегментирование»>«Настройки» на экране открывается окно настроек подсистемы, которое показано на рисунке 31.

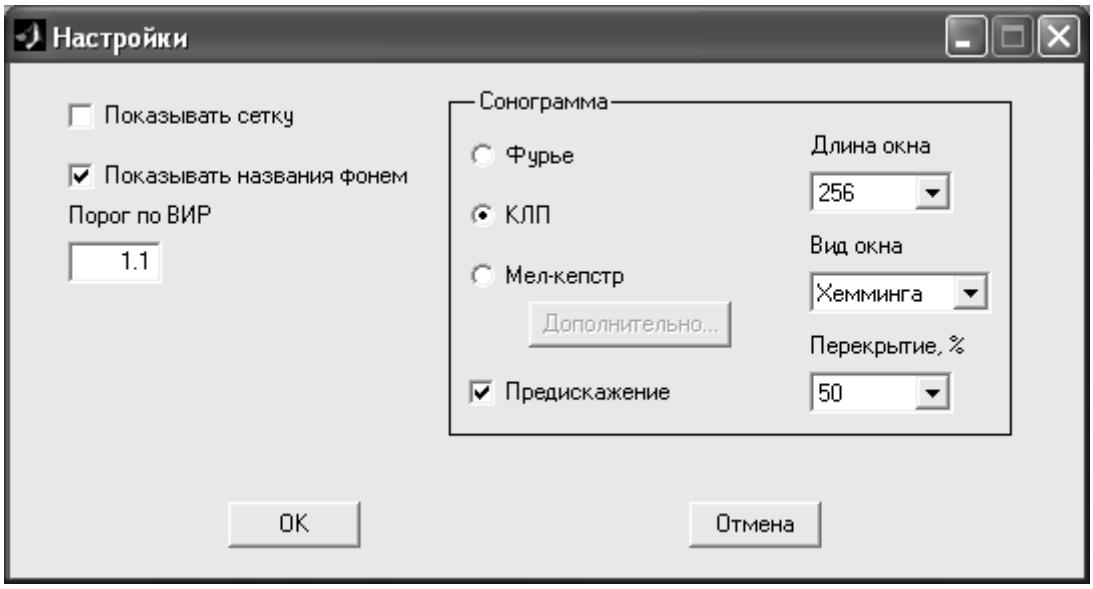



Рисунок 31. Окно настроек подсистемы

Здесь установка флажка «Показывать сетку» отображает сетку на временной диаграмме. Флажок «Показывать названия фонем» управляет отображением названий выделенных фонем. Порог по ВИР определяет максимальную величину между эталонной и выделенной фонемами, при которой принимается решение, что данная реализация фонемы соответствует какому-либо эталону выбранного диктора. Раздел «Сонограмма» определяет параметры построения сонограммы

исследуемого речевого сигнала. На рисунке 32 показан пример её построения с использованием преобразования Фурье.

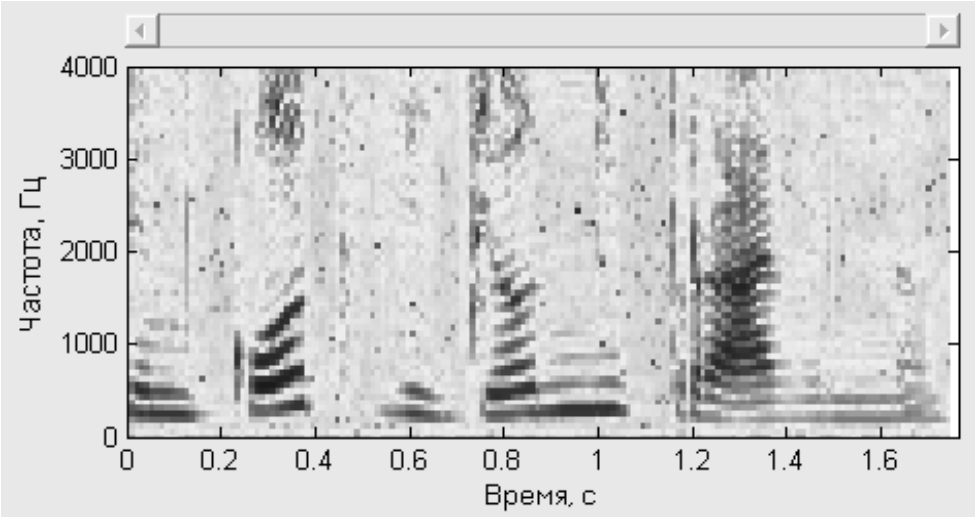



Рисунок 32. Пример построения сонограммы исследуемого речевого сигнала

Флажок «Предыскажение» определяет, будет ли при построении использоваться фильтр верхних частот с передаточной характеристикой $H(z)=1-0{,}97z^{-1}$ для компенсации влияния губ и ротовой полости на форму спектра речевого сигнала.

*Спектральный анализ фонем*

После выполнения процедуры сегментирования появляется возможность просматривать спектр выбранной фонемы в различных режимах. Существует два режима просмотра спектра:

- авторегрессионный спектр;
- Фурье-спектр.

Спектр сигнала отображается на форме «Спектр выделенного участка», которая содержит следующие меню:

- «График» содержит меню для вывода полученной диаграммы СПМ на печать или в файл в виде рисунка.
- «Сохранить» содержит следующие подменю:
  - «В файл...» позволяет сохранить график спектра в файлы на диске. Поддерживаются следующие форматы картинок:

PostScript, BMP, TIFF, JPEG, PNG, Windows Metafile, Portable Pixmap, PCX.

- «В буфер обмена» копирует текущий спектр в виде картинки в буфер обмена.

– «Установка страницы» открывает стандартный диалог Windows для установки параметров печати.
– «Печать...» отправляет график на печать в соответствии с установленными параметрами.
– «Опции» содержит меню для установки параметров отображения спектра.
– «Частотный диапазон» определяет границы частот, в которых будет строится спектр. Можно выбрать одну из следующих:
  - «По умолчанию» – в этом случае устанавливаются частотные границы от 0 Гц до половины частоты дискретизации сигнала.
  - «0 – 4 кГц» устанавливает границы по частоте от 0 до 4 кГц.
  - «0 – 3,2 кГц» устанавливает границы по частоте от 0 до 3,2 кГц.
– «АР анализ» содержит меню, устанавливающие различные режимы построения спектра по АР-модели выбранной фонемы.
  - «Предыскажение» – установка этого меню определяет, будет ли применяться к сигналу фильтр с передаточной характеристикой $H(z)=1-0{,}97z^{-1}$ перед расчётом его спектра.
  - «Окно» позволяет выбрать вид оконной функции, на которую умножается входной сигнал. В настоящий момент доступно для выбора два типа окон:
    1. Окно Хемминга.
    2. Прямоугольное окно.

- «Сохранить коэффициенты» сохраняет коэффициенты в файле в формате *csv*.

– «БПФ анализ» содержит меню, устанавливающие различные режимы построения Фурье-спектра выбранной фонемы.

- «Вид спектра» содержит меню для выбора вида отображаемого спектра, который может быть двух следующих видов:
  1. Гладкий.
  2. Линейчатый.

Чтобы выбрать интересующую фонему, нужно щелкнуть левой кнопкой мыши по соответствующему выделенному интервалу, который будет в этом случае заключен в прямоугольную рамку. На рисунке 33 показан спектр АР-модели выбранной фонемы.

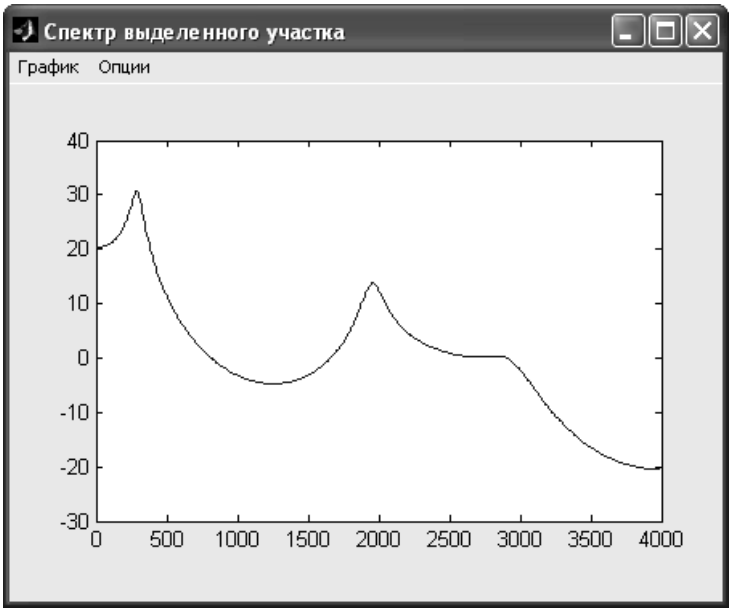



Рисунок 33. Спектр АР-модели выбранной фонемы

Управление видом спектра осуществляется на форме «Управление отображением», которая показана на рисунке 34.

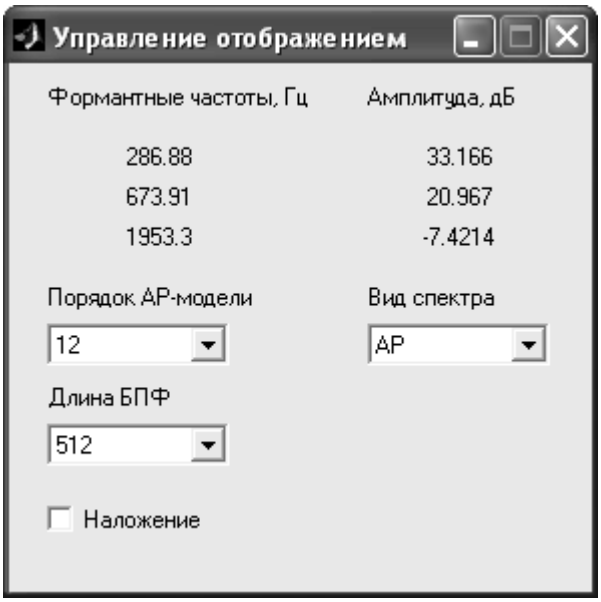



Рисунок 34. Форма «Управление отображением»

На данной форме содержатся основные элементы управления видом спектра, а также отображаются значения первых трёх формантных частот и их амплитуд. Порядок АР-модели может быть выбран в интервале 8…32. Длина БПФ выбирается в диапазоне от 64 до 1024 отсчётов. Список «Вид спектра» позволяет выбрать для отображения либо спектр АР-модели соответствующего порядка, либо спектр Фурье. На рисунке 35 показан Фурье-спектр фонемы при применении окна Хемминга.

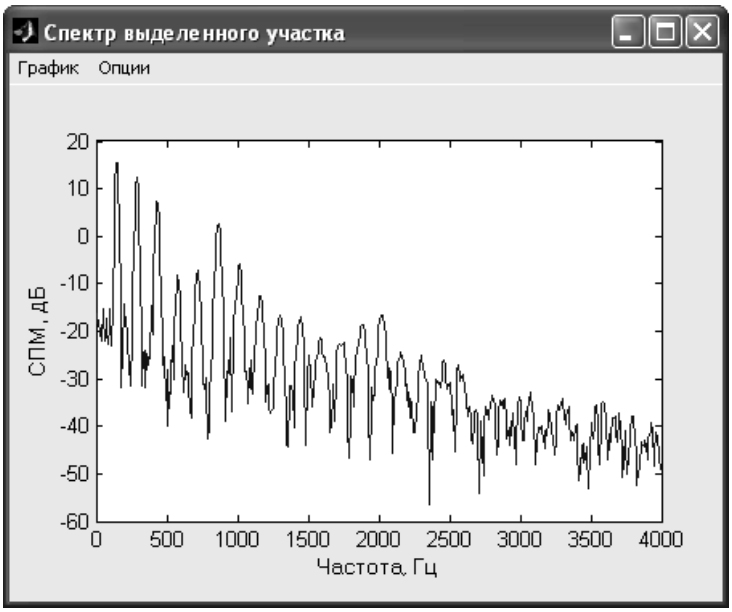



Рисунок 35. Фурье-спектр фонемы при применении окна Хемминга

Установка флажка «Наложение» приводит к тому, что выводимые спектры будут накладываться друг на друга. На рисунке 36 показан пример совмещения двух видов спектра для одной и той же фонемы.

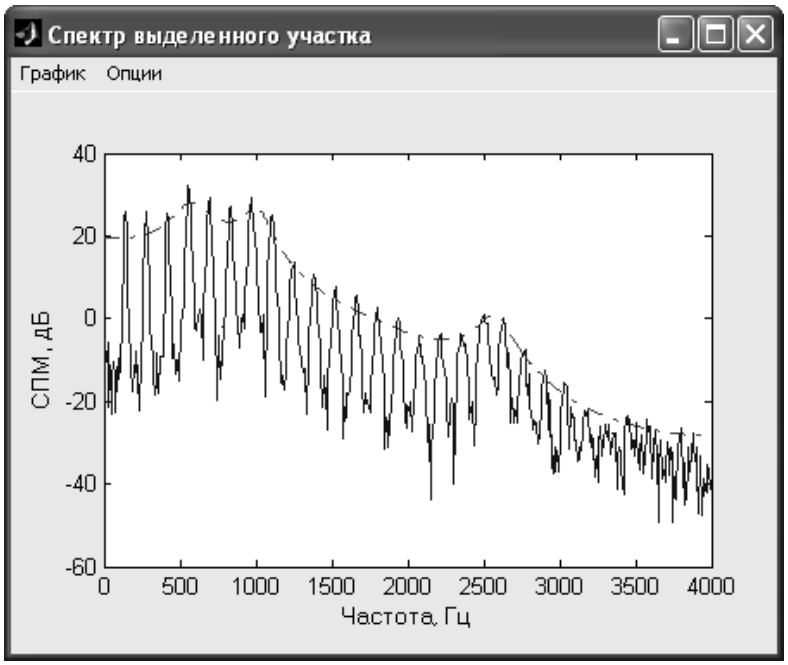



Рисунок 36. Пример совмещения двух видов спектра для одной фонемы

Если впоследствии снять флажок, то на графике будет отображаться только последний выведенный спектр.

### 3.4. Примеры применения

Рассмотренная ИС обладает уникальным спектром своих функциональных возможностей, вытекающих из новых возможностей **ИТВР**, а также 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 неопределённости, и сжатия данных, и вычислительной сложности, и главное, проблемы обновления речевых баз данных без разрушения их структуры. Рассмотрим актуальные примеры применения ИС.

*1. Автоматическая классификация устной речи по её фонетическому составу на основе критерия МИР.*

Существуют различные методы классификации дикторов. Например, по языковым группам или по местам проживания. Однако данные критерии не имеют под собой чёткого теоретического обоснования, поскольку даже в пределах одного региона фонетический состав речи разных дикторов может сильно варьироваться.

Разработанный в результате проведённых исследований математический аппарат позволяет проводить строгую классификацию дикторов по количеству выделенных из их речи фонем. Полученные результаты могут использоваться, например, при построении адаптивных систем распознавания речи.

*2. Автоматический анализ качества устной речи по её звуковому ряду на основе критерия МИР.*

В данном случае под качеством устной речи понимается количество выделенных фонем из текста, произнесенного диктором. Т. е. чем больше их число, тем более богатой в фонетическом смысле является речь данного диктора. Соответственно, если у первого диктора было выделено больше фонем, чем у второго, при тех же самых параметрах, то произношение первого диктора можно считать более качественным.

Разработанная система позволяет проводить данный анализ. В результате можно получить не только количество фонем, но и сами эти фонемы.

*3. Автоматическая постановка правильного произношения по критерию МИР.*

В настоящее время большинство систем автоматической постановки произношения основаны на сопоставлении произнесенного слова или фонемы с заранее подготовленным эталоном. При этом чем меньше

расстояние между эталоном и входным сигналом, тем лучше считается произношение.

Недостатком данного подхода является то, что диктор может быть не в состоянии произнести звук в достаточной степени похожий на эталон. Выходом из данной ситуации может стать сравнение произношения с несколькими эталонами. Диктору будет достаточно приблизить свое произношение к одному из них. В данной ситуации имеет смысл говорить не об абсолютной, а об относительной близости произношения к какому-то из эталонов. Это позволяет сократить время на обучение, поскольку диктору будет достаточно приблизить свое произношение к наиболее удобному для него варианту.

На основе разработанной системы можно легко построить программное обеспечение, реализующее вышеизложенный принцип. Подобные программы могут использоваться при обучении иностранным языкам. Также их можно применять при обучении глухих людей и в других областях, где требуется постановка аппроксимированного произношения.

## 4. Задания к лабораторному практикуму

Объектом исследований в рамках лабораторного практикума служит рассмотренная выше (гл. 3) ИС ФАР, реализованная в виде компьютерной программы на базе MATLAB-7.2. Студентам предлагается следующий план лабораторных работ:

1. Изучение работы ИС и тестирование её базовых функций.

Результаты работы оформляются студентами в виде Протокола. Его форма дана в приложении 1.

2. Применение ИС для решения ряда актуальных задач:
   a. *Автоматическая классификация устной речи по её фонетическому составу на основе критерия МИР.*
   b. *Автоматический анализ качества устной речи по её звуковому ряду на основе критерия МИР.*
   c. *Автоматическая постановка правильного произношения по критерию МИР.*

Результаты работы оформляются студентами также в виде протокола. Его примерная форма дана в приложении 2.

**Библиографический список**


1. Савченко В.В. Информационная теория восприятия речи // Изв. вузов России. Радиоэлектроника. 2007. Вып. 6. С. 3-9.
2. Савченко В.В., Васильев Р.А. Анализ эмоционального состояния диктора по голосу на основе фонетического детектора лжи // Научные ведомости БелГУ. Серия: История. Политология. Экономика. Информатика. 2014. № 21 (192). Вып. 32/1. С. 186-195.
3. Савченко В.В. Определение объёма контрольной выборки в условиях априорной неопределенности по принципу гарантированного результата // Научные ведомости БелГУ. Серия: История. Политология. Экономика. Информатика. 2015. № 1(198). Вып. 33/1. С.74-78.
4. Савченко В.В., Акатьев Д.Ю. Теоретико-информационное обоснование МОФ в задачах автоматического распознавания речи // Сист. упр-я и информационные технологии. 2008. №1 (31). С. 60-66.
5. Савченко В.В. Теория вероятностей и математическая статистика: Конспект лекций. Н. Новгород: НГЛУ, 2007.
6. Акатьев Д.Ю., Чикова Н.Д. Сборник задач по теории вероятности и математической статистике. Н. Новгород: НГЛУ, 2007.
7. Савченко В.В., Никольская В.А. Математика и информатика для студентов лингвистических специальностей. Н. Новгород: НГЛУ, 2008.
8. Савченко В.В., Грушин В.А. Актуальные главы высшей математики. Конспект лекций и планы семинарских занятий. Н. Новгород: НГЛУ, 2003.
9. Савченко В.В., Акатьев Д.Ю. Исследование коммуникативных особенностей русского языка на основе теоретико-информационного подхода // Сборник трудов VII международной научно-практической конференции по региональной культуре. Н. Новгород: НГЛУ, 2006. С.121-122.

10. Савченко В.В., Акатьев Д.Ю. Автотестирование качества произношения речи по принципу минимального информационного рассогласования // Современные тенденции компьютеризации процесса изучения иностранных языков: Сборник научных трудов. Луганск: Восточно-украинский национальный университет. 2005. Вып. 3. С. 205-206.

**Приложение 1**

**Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования**
**«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ЛИНГВИСТИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. Н.А. ДОБРОЛЮБОВА»**

---

# Протокол испытаний информационной системы фонетического анализа речи

***ФОРМА – ОБРАЗЕЦ ОФОРМЛЕНИЯ***

**Руководитель НИР**

д.т.н. проф. Савченко В.В.

---

**Нижний Новгород – 2015**

1. ***Объект исследований:*** информационная система фонетического анализа слитной речи (ИС), реализованная в виде компьютерной программы (системы) с применением MATLAB-7.2.

Алгоритмы обработки речевых сигналов, примененные в данной системе, базируются на информационной теории восприятия речи.

2. ***Время проведения исследований:*** период с _____по _____ 20__ г.

3. ***Программа экспериментальных исследований.*** На первом этапе была проверена работоспособность базовых функций системы, таких как: корректность процедуры загрузки и завершения работы, правильность работы с дикторами и группами дикторов (создание, удаление, изменение).

На втором этапе проводилась проверка работы алгоритмов выделения фонем из входных данных в различных режимах работы ИС. Для тестирования был выбран текст объёмом 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 применялись 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. Из полученных звуковых файлов производилось выделение фонем и запись полученных результатов в базу данных.

На третьем этапе проводилось тестирование работоспособности функций визуализации результатов, а на четвертом – тестирование подсистемы автоматического транскрибирования речи.

Тестирование проводилось на персональном компьютере следующей конфигурации: Pentium IV, 256 Мбайт ОЗУ, Windows XP, Matlab 6.5.

4. ***Полученные результаты*** иллюстрируются следующей таблицей данных. Здесь символом ‘+’ обозначены успешно пройденные тесты, а символом ‘-’ – непройденные. Перед каждым разделом приведены рисунки, на которых показаны формы, которые были задействованы в том или ином виде тестирования.

## Тестирование базовых функций

Для запуска приложения необходимо запустить среду разработки *Matlab*, интерфейс которой показан на рисунке 1.

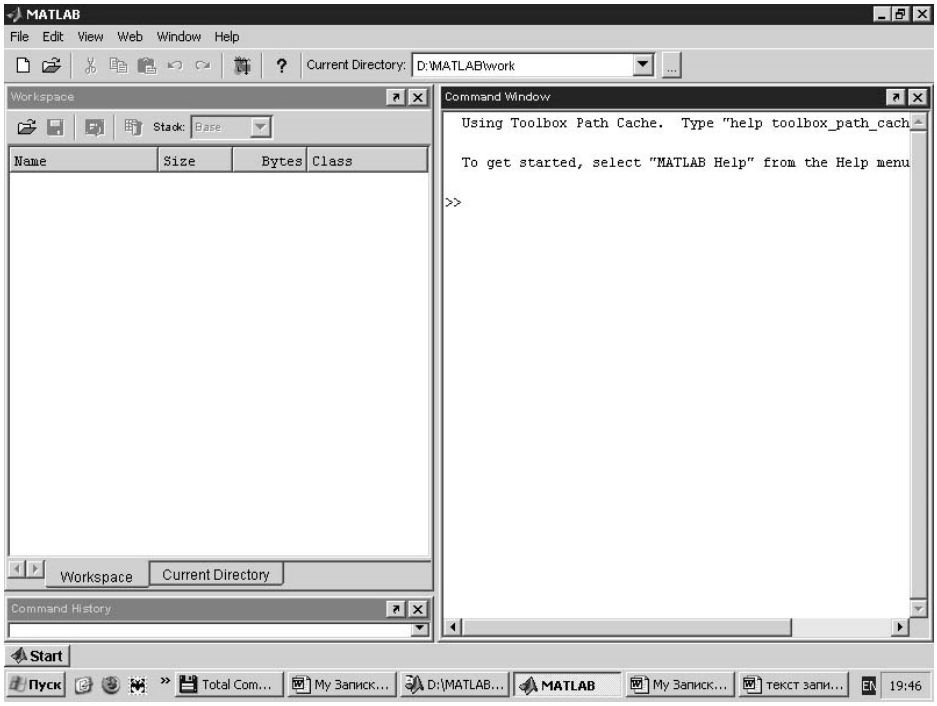



Рисунок 1. Интерфейс среды разработки *Matlab*

На рисунке 2 изображено главное окно программы.

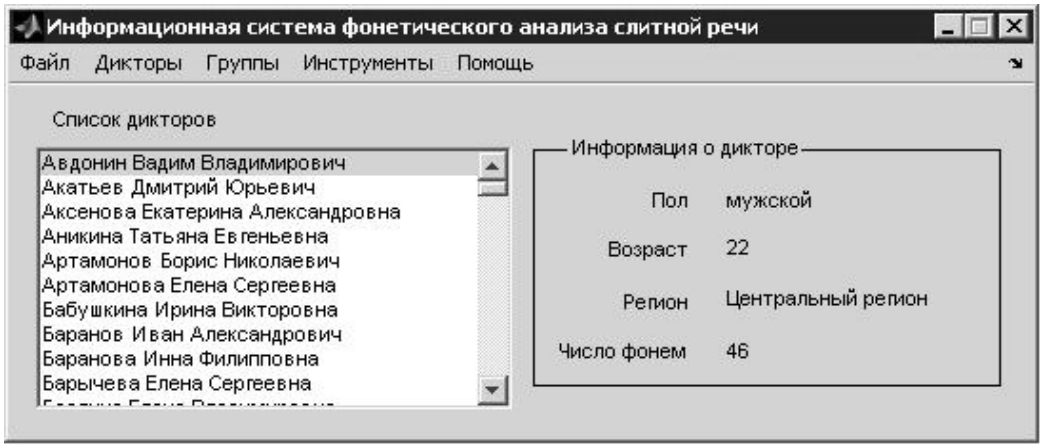



Рисунок 2. Главное окно программы

На этой форме отображаются дикторы, внесенные в БД и главное меню программы. При выборе любого диктора из списка в правой части окна отображается краткая информация о нём.

На рисунке 3 показана форма ввода информации о дикторе.

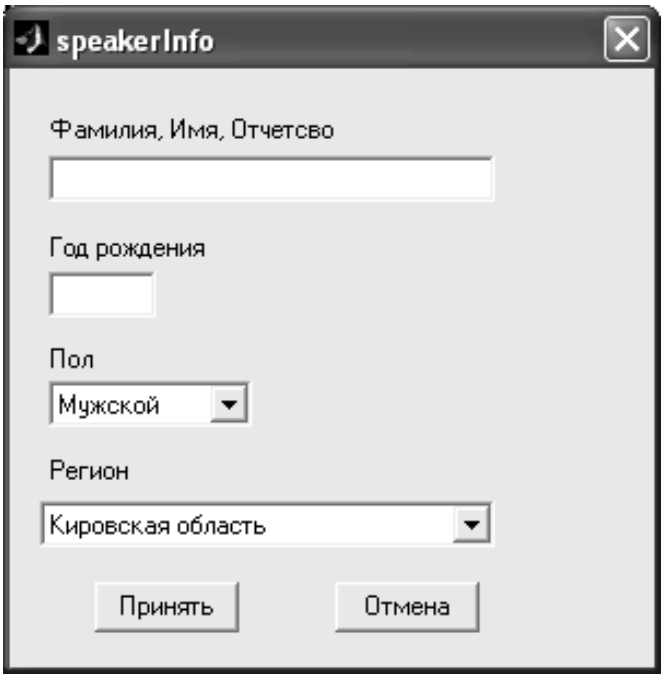



Рисунок 3. Форма ввода информации о дикторе

На рисунке 4 показана форма работы с группами дикторов.

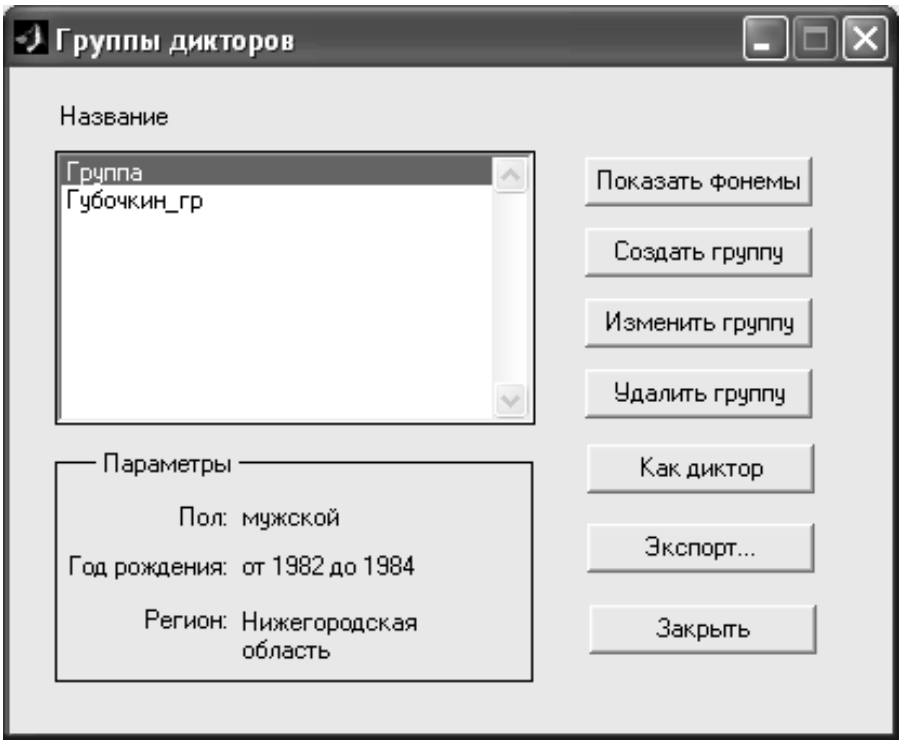



Рисунок 4. Форма работы с группами дикторов

На рисунке 5 показано окно основных настроек информационной системы.

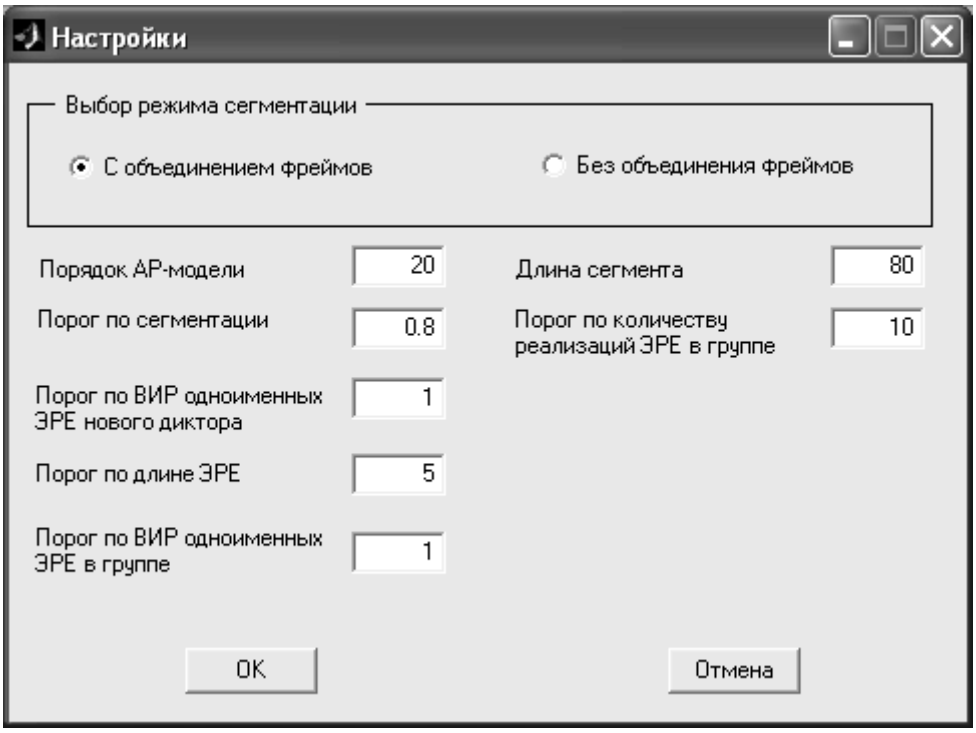



Рисунок 5. Окно основных настроек информационной системы

Таблица 1

| **Название теста** | **Результат выполнения** | **Примечание** |
|---|---|---|
| Проверка корректности работы процесса загрузки информационной системы | + | Для запуска программы необходимо ввести в командной строке системы *matlab* команду *phonemexplorer* и нажать клавишу ввода. В результате на экране появится главное окно программы. Следует отметить, что для корректного выполнения в поле *Current Directory* следует указать путь к каталогу с программой (например, “D:\MATLAB\work”). |

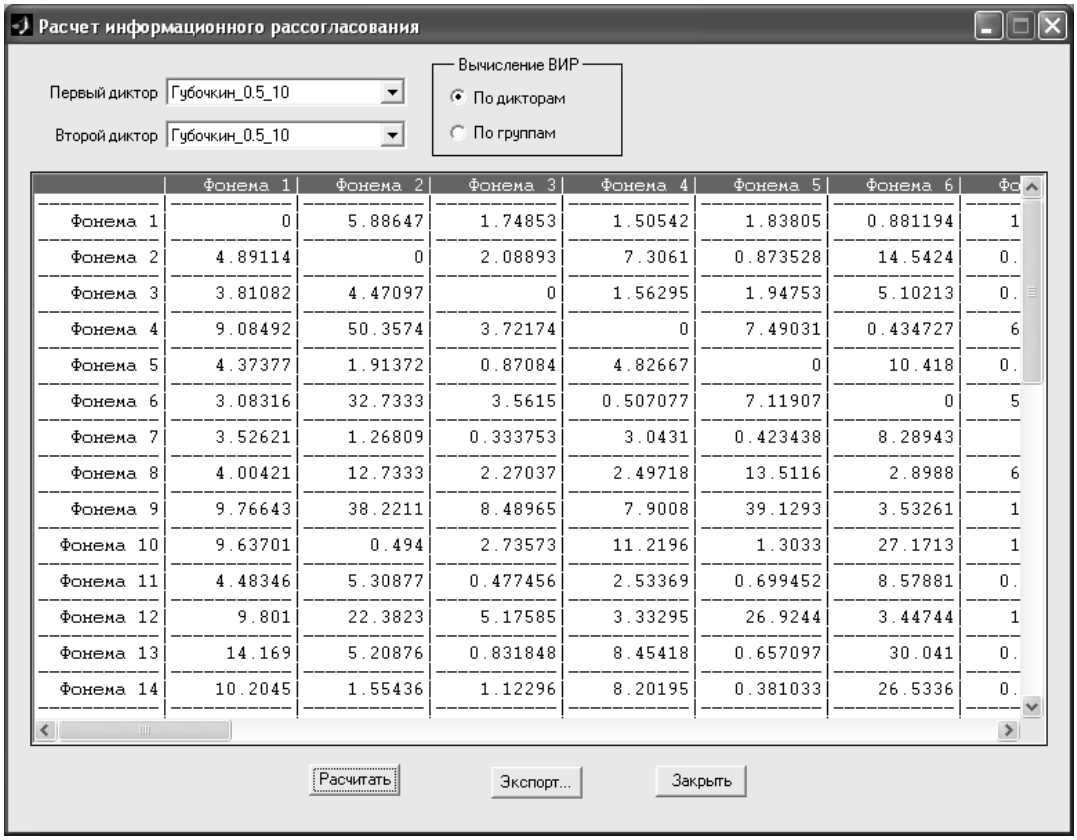
*Продолжение таблицы 1 на с. 75.*

*Продолжение таблицы 1.*

| | | |
|---|---|---|
| Проверка корректности завершения работы информационной системы | + | |
| Проверка корректности правильности добавления нового диктора в систему | + | |
| Проверка корректности правильности удаления выбранного диктора из системы | + | |
| Проверка корректности изменения информации о существующем дикторе | + | |
| Проверка создания новой группы с заданными критериями отбора | + | |
| Проверка корректности изменения критериев отбора существующей группы | + | |
| Проверка корректности удаления существующей группы | + | |
| Проверка корректности сохранения настроек программы | + | |

**Тестирование функций выделения и обработки фонем**

На рисунке 6 показано окно расчета матрицы ВИР между двумя дикторами или группами дикторов.




| | Фонема 1 | Фонема 2 | Фонема 3 | Фонема 4 | Фонема 5 | Фонема 6 | Фо |
|---|---|---|---|---|---|---|---|
| Фонема 1 | 0 | 5.88647 | 1.74853 | 1.50542 | 1.83805 | 0.881194 | 1 |
| Фонема 2 | 4.89114 | 0 | 2.08893 | 7.3061 | 0.873528 | 14.5424 | 0. |
| Фонема 3 | 3.81082 | 4.47097 | 0 | 1.56295 | 1.94753 | 5.10213 | 0. |
| Фонема 4 | 9.08492 | 50.3574 | 3.72174 | 0 | 7.49031 | 0.434727 | 6 |
| Фонема 5 | 4.37377 | 1.91372 | 0.87084 | 4.82667 | 0 | 10.418 | 0. |
| Фонема 6 | 3.08316 | 32.7333 | 3.5615 | 0.507077 | 7.11907 | 0 | 5 |
| Фонема 7 | 3.52621 | 1.26809 | 0.333753 | 3.0431 | 0.423438 | 8.28943 | |
| Фонема 8 | 4.00421 | 12.7333 | 2.27037 | 2.49718 | 13.5116 | 2.8988 | 6 |
| Фонема 9 | 9.76643 | 38.2211 | 8.48965 | 7.9008 | 39.1293 | 3.53261 | 1 |
| Фонема 10 | 9.63701 | 0.494 | 2.73573 | 11.2196 | 1.3033 | 27.1713 | 1 |
| Фонема 11 | 4.48346 | 5.30877 | 0.477456 | 2.53369 | 0.699452 | 8.57881 | 0. |
| Фонема 12 | 9.801 | 22.3823 | 5.17585 | 3.33295 | 26.9244 | 3.44744 | 1 |
| Фонема 13 | 14.169 | 5.20876 | 0.831848 | 8.45418 | 0.657097 | 30.041 | 0. |
| Фонема 14 | 10.2045 | 1.55436 | 1.12296 | 8.20195 | 0.381033 | 26.5336 | 0. |

Рисунок 6. Окно расчета матрицы ВИР между двумя дикторами или группами дикторов

Таблица 2

| **Название теста** | **Результат выполнения** | **Примечание** |
|---|---|---|
| Проверка выделения фонем для выбранного диктора из звукового файла | + | Было проведено выделение фонем из звукового файла для одного диктора при разных параметрах и для двух дикторов при одинаковых параметрах. В первом случае было выделено 75 и 69 фонем, а во втором 73 и 57. |

*Продолжение таблицы 2 на с. 77.*

*Продолжение таблицы 2.*

| | | |
|---|---|---|
| Проверка выделения фонем для выбранного диктора из группы звуковых файлов | + | |
| Проверка добавления фонем для выбранного диктора из звукового файла | + | |
| Проверка удаления фонем выбранного диктора | + | |
| Проверка корректности экспорта АР-коэффициентов фонем выбранного диктора в текстовый файл | + | |
| Проверка корректности визуализации фонем выбранного диктора | + | |
| Проверка формирования группы по выбранным дикторам | + | Было сформировано две группы. Первая группа формировалась из фонем двух различных дикторов. При этом было выделено 110 фонем. Вторая группа формировалась из фонем одного диктора, выделенных при различных параметрах обработки. Данная группа состояла из 86 фонем. Порог по ВИР одноименных ЭРЕ в группе в обоих случаях был фиксированным и равным 1,1. |
| Проверка корректности формирования группы по заданным критериям отбора | + | |

*Продолжение таблицы 2 на с. 78.*

*Продолжение таблицы 2.*

| | | |
|---|---|---|
| Проверка сохранения фонем выбранной группы как фонем диктора | + | |
| Проверка корректности экспорта фонем выбранной группы в текстовый файл | + | |
| Проверка корректности визуализации фонем выбранной группы | + | |
| Проверка корректности расчета матрицы ВИР между двумя дикторами | + | |
| Проверка корректности расчета матрицы ВИР между двумя группами | + | |
| Проверка корректности экспорта матрицы ВИР в текстовый файл | + | |

## Тестирование функций визуализации и воспроизведения фонем

На рисунке 7 показана форма визуализации и воспроизведения фонем.

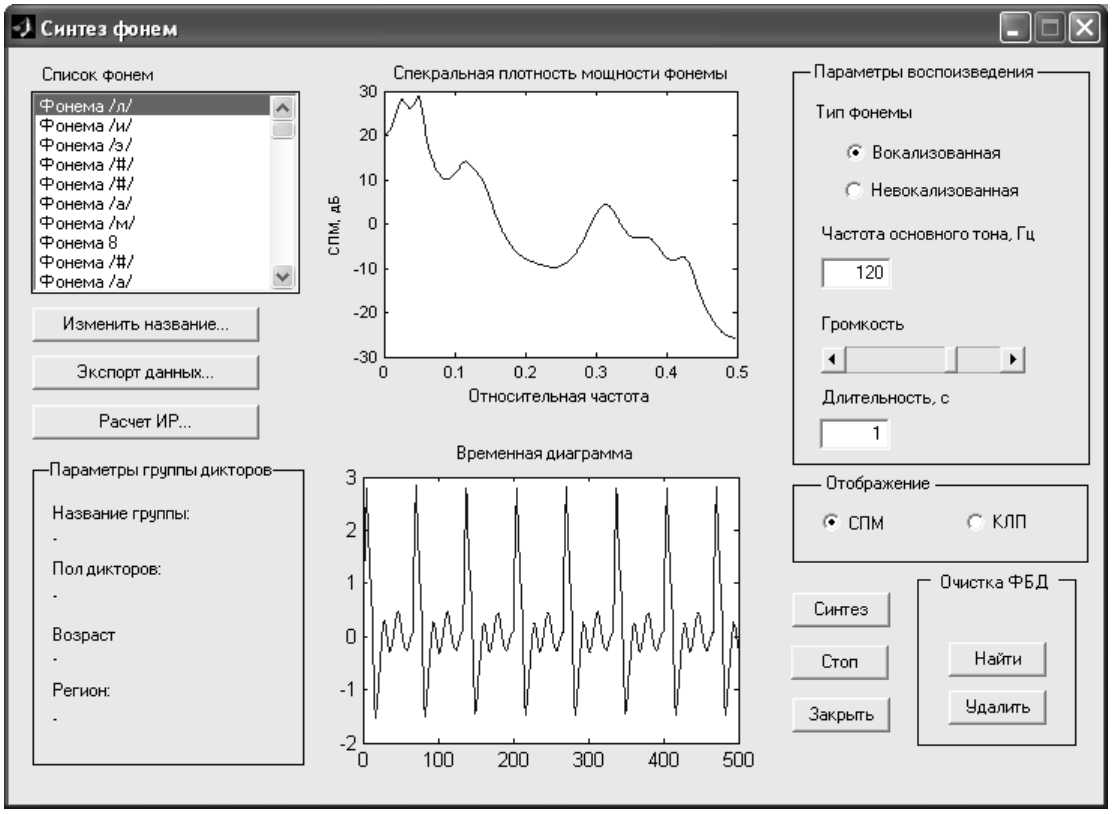
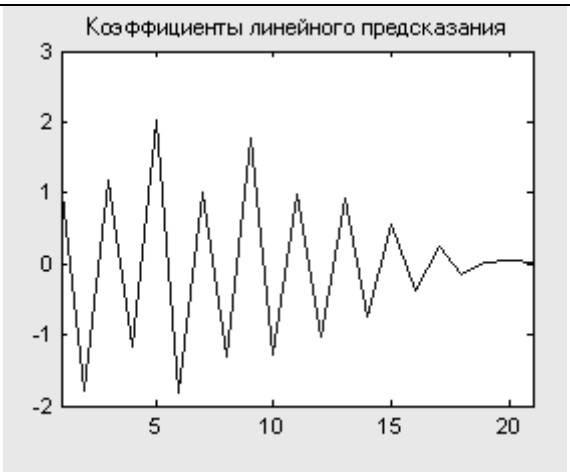
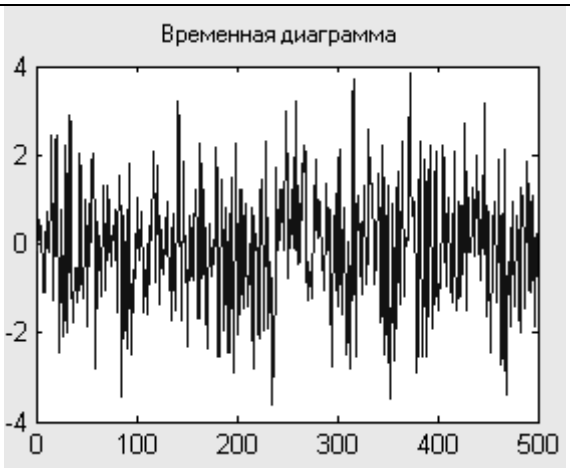


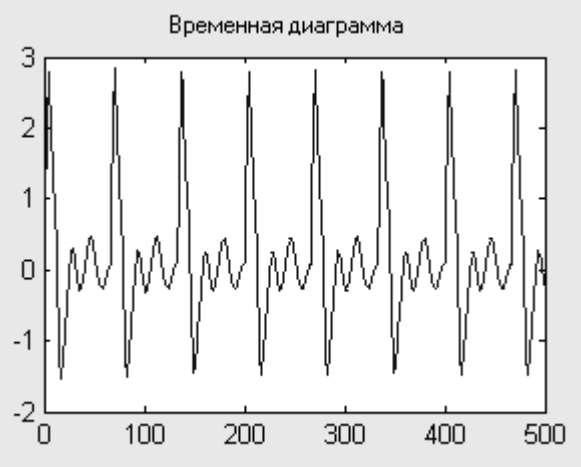

Рисунок 7. Форма визуализации и воспроизведения фонем

Таблица 3

| **Название теста** | **Результат выполнения** | **Примечание** |
|---|---|---|
| Проверка корректности построения графика СПМ для выбранной фонемы | + | |
| Проверка корректности вывода значение КЛП для выбранной фонемы | |  |
| Проверка корректности построения временной диаграммы для невокализованной фонемы | + |  |

*Продолжение таблицы 3 на с. 81.*

*Продолжение таблицы 3.*

| | | |
|---|---|---|
| Проверка корректности построения временной диаграммы для вокализованной фонемы при заданном значении основного тона | + |  |
| Проверка воспроизведения вокализованной фонемы заданной длительности | + | |
| Проверка воспроизведения невокализованной фонемы заданной длительности | + | |
| Проверка изменения названия выбранной фонемы | + | |
| Проверка экспорта АР-коэффициентов выбранной фонемы в текстовый файл | + | |
| Проверка расчета значений ВИР между выбранной фонемой и набором звуковых файлов | + | |

## Тестирование функции автоматического транскрибирования речи

На рисунке 8 показана главная форма подсистемы автоматического транскрибирования.

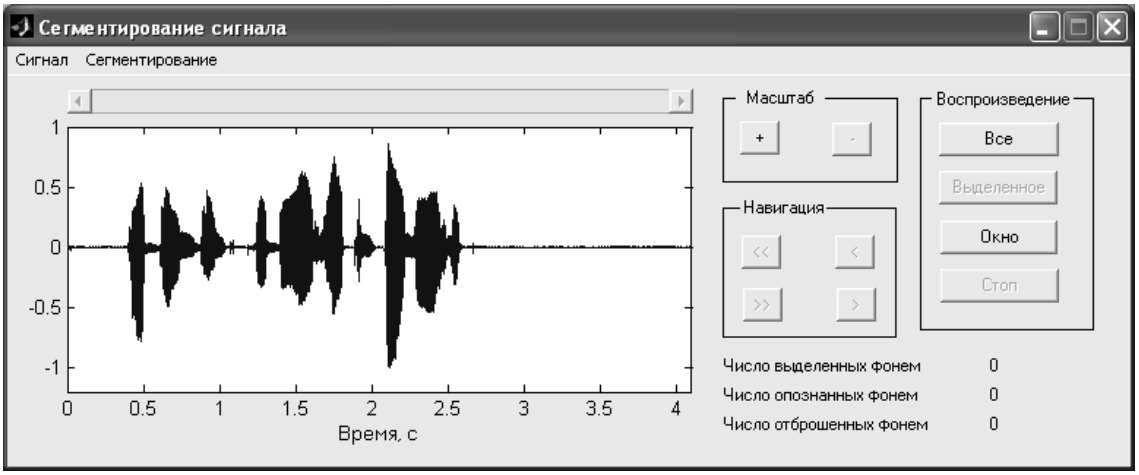



Рисунок 8. Главная форма подсистемы автоматического транскрибирования

На рисунке 9 показано окно настроек подсистемы.

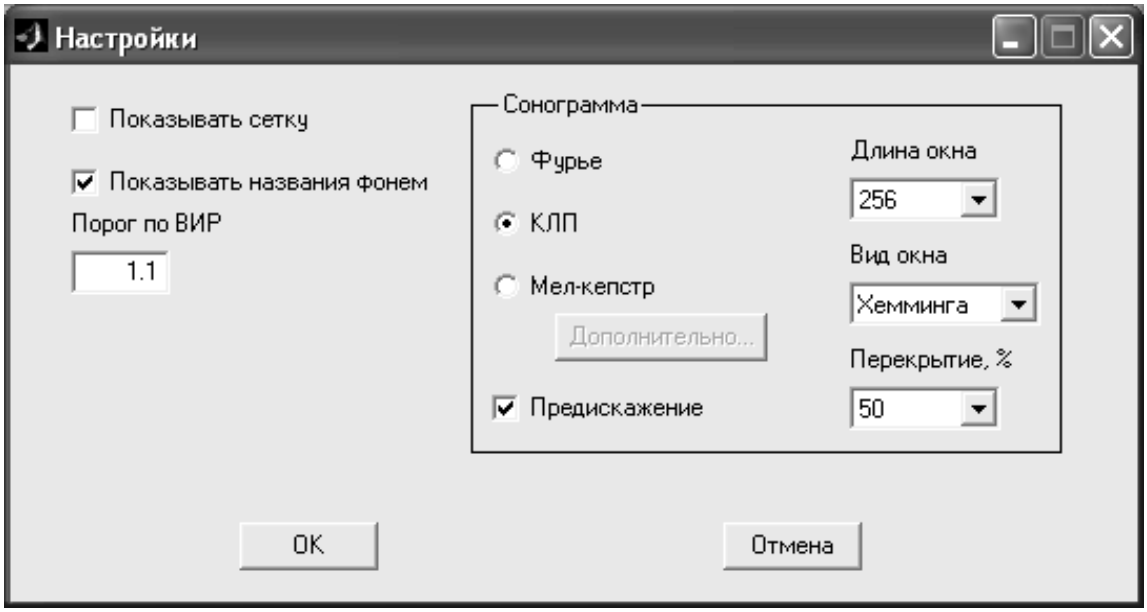



Рисунок 9. Окно настроек подсистемы

На рисунке 10 показано окно отображения спектра выделенного фрагмента сигнала.

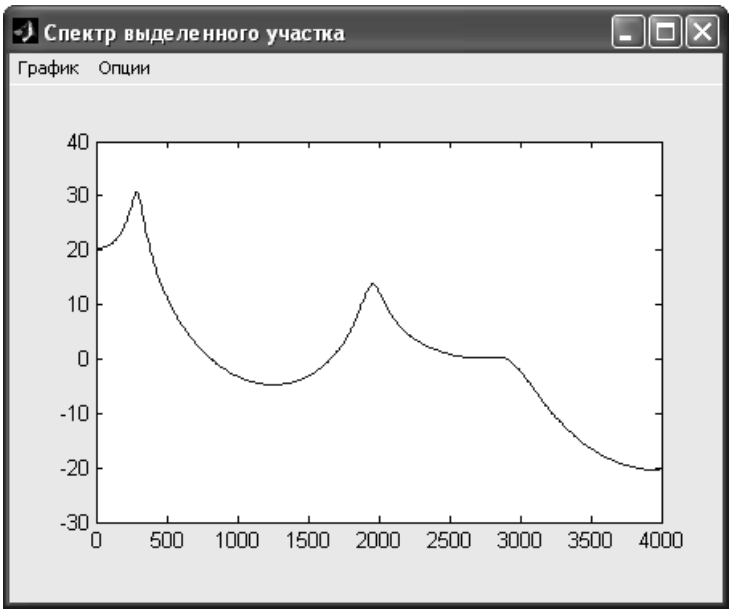



Рисунок 10. Окно отображения спектра выделенного фрагмента сигнала

Управление видом спектра осуществляется на форме «Управление отображением», которая показана на рисунке 11.

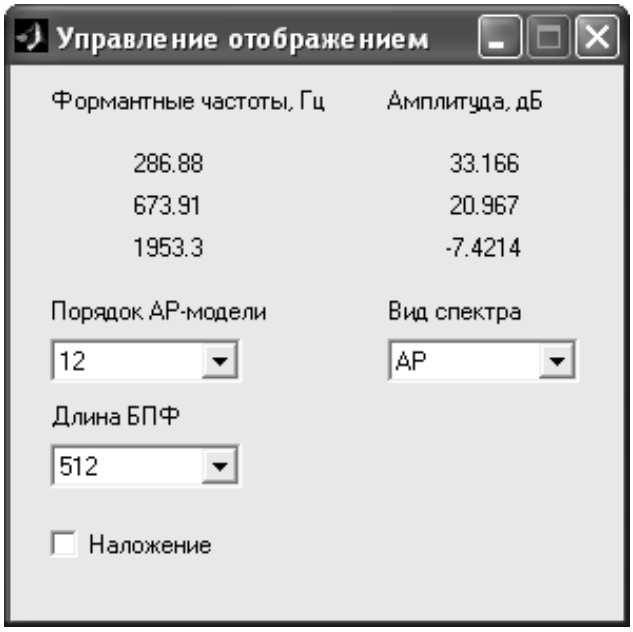
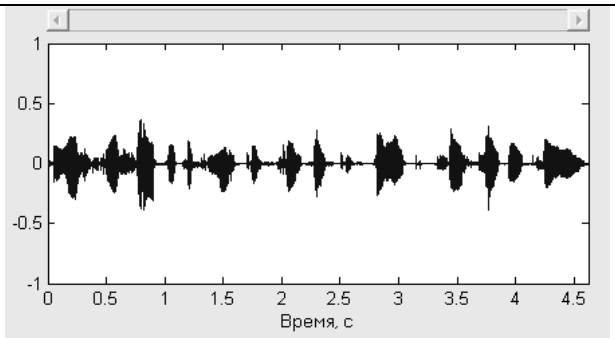
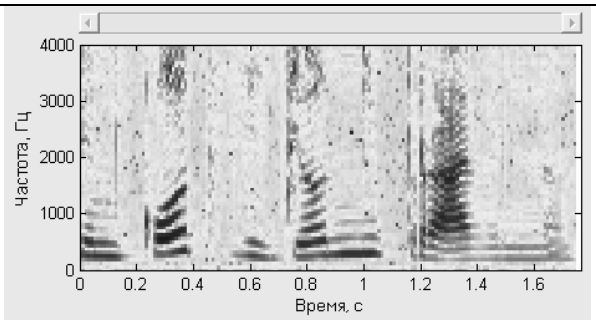


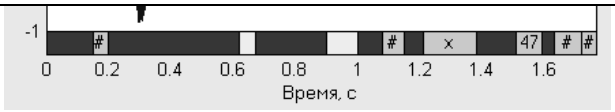
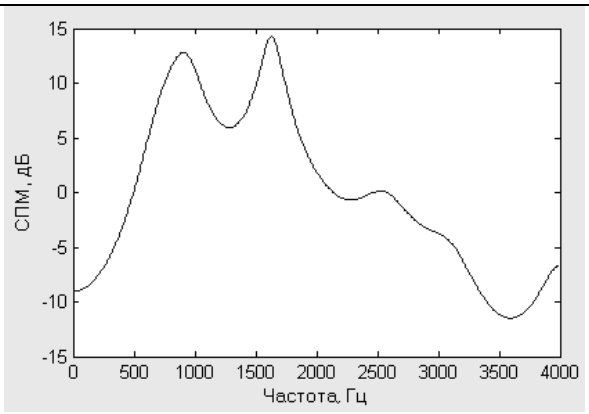
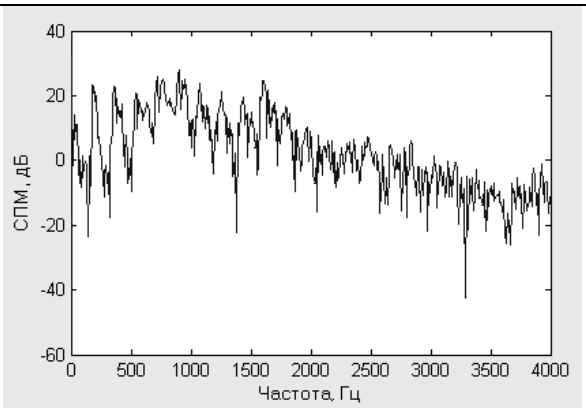

Рисунок 11. Форма «Управление отображением»

Таблица 4

| **Название теста** | **Результат выполнения** | **Примечание** |
|---|---|---|
| Проверка загрузки звукового файла | + |  |
| Проверка сохранения полученного графика в графические файлы различных форматов | + | |
| Проверка установки страницы для печати временной диаграммы | + | |
| Проверка вывода временной диаграммы сигнала на печать | + | |
| Проверка корректности сохранения настроек подсистемы АТР | + | |
| Проверка отображения в различных режимах сонограммы загруженного сигнала | + |  |

*Продолжение таблицы 4 на с. 85.*

*Продолжение таблицы 4.*

| | | |
|---|---|---|
| Проверка результатов работы алгоритма АТР | + |  |
| Проверка построения графика СПМ, рассчитанной по АР-коэффициента выделенного фрагмента сигнала при различных порядках АР-модели | + |  |
| Проверка определения формантных частот выделенного фрагмента сигнала | + | |
| Проверка построения графика СПМ, выделенного фрагмента сигнала при различных порядках БПФ | + |  |
| Проверка воспроизведения загруженного речевого сигнала | + | |
| Проверка воспроизведения выделенного фрагмента речевого сигнала | + | |

*Продолжение таблицы 4 на с. 86.*

*Продолжение таблицы 4.*

| | | |
|---|---|---|
| Проверка воспроизведения фрагмента загруженного сигнала, отображаемого на графике | + | |
| Проверка увеличения масштаба отображаемого речевого сигнала | + | |
| Проверка уменьшения масштаба отображаемого речевого сигнала | + | |
| Проверка работоспособности навигации по анализируемому сигналу | + | |
| Проверка корректности отображения названий выделенных фонем | + | |

## ВЫВОДЫ

Рассмотренная ИС обеспечивает выполнение следующих функций:

- автоматическое выделение фонем из входных данных;
- обработка фонем;
- хранение фонем;
- визуализация полученных результатов.

В процессе тестирования перечисленных функций выявлены следующие дефекты:

---

---

---

**Исполнители:** к.т.н. доц. Акатьев Д.Ю.,

инженер Губочкин И.В.

**Приложение 2**

**Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования**
**«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ЛИНГВИСТИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. Н.А. ДОБРОЛЮБОВА»**

Факультет международных отношений, экономики и управления

Кафедра математики и информатики

Лаборатория речевых технологий

**ОТЧЕТ**
**к лабораторной работе**

по курсу «Информационная теория восприятия речи»
ООП 090900.62 - Информационная безопасность

---

на тему **«ИНФОРМАЦИОННАЯ СИСТЕМА ФОНЕТИЧЕСКОГО АНАЛИЗА РЕЧИ»**

ЗАДАНИЕ №_1_: *Автоматическая классификация устной речи по ее фонетическому составу на основе критерия МИР*

**Работу сдал** _____
Группа _____

**Работу принял**_____
Кафедра МИ

**Нижний Новгород – 201_**

## 1. ЗАДАНИЕ К РАБОТЕ

_____
_____
_____

## 2. *ВРЕМЯ ПРОВЕДЕНИЯ ИССЛЕДОВАНИЙ*

период с _____

## 3. *ПРИБОРЫ И ОБОРУДОВАНИЕ:*

- ноутбук _____ с установленной программой ИС ФАР (см. приложение 1);
- встроенный в ноутбук микрофон;
- встроенная в ноутбук акустическая система.

## *4. ПРОГРАММА ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ*

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

***СОСТАВ КОНТРОЛЬНОЙ ГРУППЫ ДИКТОРОВ (ФИО, группа, возраст)*:**

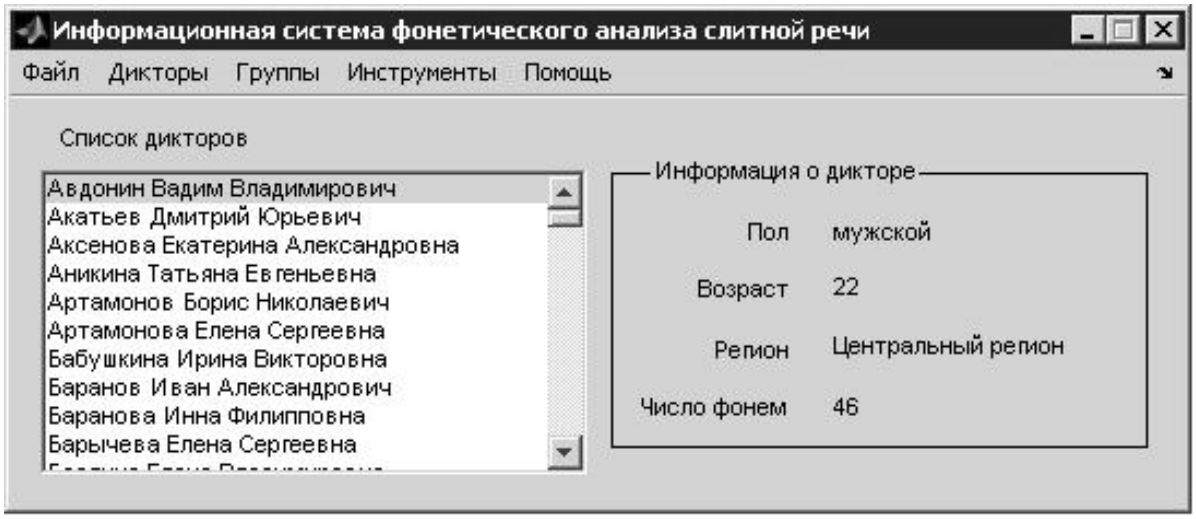
**1)**_____

**2)** _____

**3)** _____

## *5. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ*

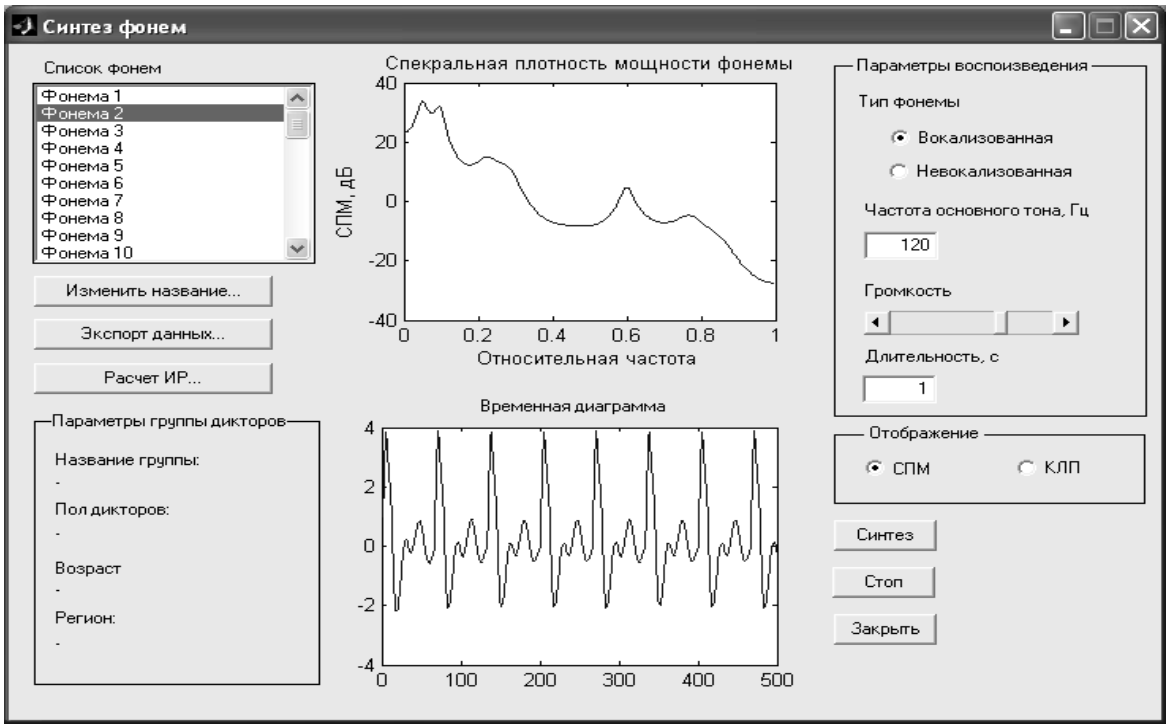
А) Режим настройки параметров




_____
_____
_____
_____
_____
_____

Б) Рабочий режим

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____




## *6. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ВЫВОДЫ:*

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Отчет составил

студент_____

Владимир Васильевич Савченко
Дмитрий Юрьевич Акатьев

**ИНФОРМАЦИОННАЯ ТЕОРИЯ ВОСПРИЯТИЯ РЕЧИ**

**Учебное пособие**

Редакторы: А.О. Кузнецова
Д.В. Носикова
А.С. Паршаков

Лицензия ПД № 18-0062 от 20.12.2000

---

Подписано к печати Формат 60 х 90 1/16
Печ. л. Тираж экз. Заказ
Цена договорная

---

Типография ФГБОУ ВПО «НГЛУ»
603155, Н. Новгород, ул. Минина, 31а