УДК 004.93

# ФОНЕТИКО-МОРФОЛОГИЧЕСКАЯ РАЗМЕТКА РЕЧЕВЫХ КОРПУСОВ ДЛЯ РАСПОЗНАВАНИЯ И СИНТЕЗА РУССКОЙ РЕЧИ

***А. Л. Ронжин,***
*канд. техн. наук, старший научный сотрудник*
***А. А. Карпов,***
*аспирант*
*Санкт-Петербургский институт информатики и автоматизации РАН*
***Б. М. Лобанов,***
*доктор техн. наук, заведующий лабораторией*
***Л. И. Цирульник,***
*научный сотрудник*
*Объединенный институт проблем информатики НАН Беларуси*
***О. Йокиш***
*менеджер проектов*
*Дрезденский технологический университет*

*Показаны особенности подготовки и обработки текстовых данных для разметки речевых корпусов, обсуждается процедура анализа и сопоставления речевых сигналов в процессе разметки, описаны процедуры транскрибирования русских текстов, морфемного представления языка и речи. Описываются автоматизированные системы создания фонетико-акустических речевых баз данных для персонализированного синтеза речи по тексту и распознавания русской речи SIRIUS.*

*The particularities of preparation and processing of text data for speech corpora labeling are presented. The problem of analysis and comparison of speech signals during the labeling is discussed. The procedures of Russian text transcription and morphemic description of language and speech are presented. The implementation of phonetic-morphemic labeling of speech corpora for Russian speech recognition system SIRIUS is shown. The automated system for phonetic-acoustical voice database creation for personalized speech synthesis is described.*

## Введение

Речевые технологии в России и за рубежом стремительно развиваются благодаря таким преимуществам перед типовыми средствами общения человека с машиной, как естественность, оперативность, освобождение рук и зрения пользователя, возможность управления в экстремальных условиях эксплуатации. Системы распознавания и синтеза речи уже используются в реальных приложениях в США и Европе, но в России, к сожалению, пока существуют лишь отдельные разработки. Основная проблема связана со спецификой русского языка, в частности, со сложным механизмом словообразования и фонетической интерпретации.

При построении систем распознавания и синтеза речи одной из наиболее важных задач является сегментация и маркировка баз данных (БД) речевых сигналов на минимальные семантически и фонетически значимые единицы речи. В качестве таких единиц при распознавании слитной речи целесообразно использовать морфемы, а при синтезе речи и распознавании фонем — аллофоны. Полученные сегменты сохраняются в БД и служат для обучения акустических моделей в системе распознавания речи, а также для генерации голоса клонируемого диктора в системе синтеза речи по тексту. При этом посредством ограниченного количества морфем создаются всевозможные словоформы (до нескольких миллионов), а посредством аллофонов генерируется необходимое звуковое разнообразие речевого сигнала. Одним из важнейших требований к сегментации и маркировке БД является точность разметки, от нее зависит каче-

ство синтеза и точность распознавания в конечном итоге. Обычно подготовка речевого корпуса осуществляется вручную экспертом-фонетистом с использованием полуавтоматических средств просмотра осциллограмм и спектрограмм сигнала. При высокой квалификации эксперта «ручной» метод обеспечивает достаточно точную разметку речевого корпуса, но требует много времени и усилий.

В данной работе предложена система автоматизированной обработки речевых корпусов, в основу которой положена идея «анализа через синтез» [1]. Система создана в рамках исследований по клонированию голоса и дикции личности [2, 3], созданию многоязычного и многоголосого синтезатора речи по тексту [4] и дикторонезависимой системы распознавания русской слитной речи [5].

## Выбор лингвистических единиц для задач распознавания и синтеза речи

Языки славянской группы относят к числу синтетических языков, которые характеризуются тенденцией к объединению в рамках одной словоформы лексической морфемы (или нескольких лексических морфем) и одной или нескольких грамматических морфем. Более сложная структура словообразования ведет к росту размера распознаваемого словаря, что значительно уменьшает точность и скорость распознавания. В коммерческих системах распознавания речи для английского языка (от фирм Microsoft, Dragon Systems) используется словарь свыше 100 тыс. слов. Для русского языка за счет наличия приставок, суффиксов и окончаний этот словарь возрастает более чем на порядок. При сравнении языковых моделей часто используют две характеристики: коэффициент сложности (perplexity) и число пропущенных слов или число слов, не вошедших в словарь распознавания (Out Of Vocabulary). Коэффициент сложности модели языка равен усредненному числу слов, которые могут быть связаны с предыдущим словом во фразе. В табл. 1 представлена сравнительная характеристика русского и английского языков по этим параметрам [6].

■ ***Таблица 1.*** *Сравнение статистических моделей английского и русского языков*

| Размер словаря, тыс. слов | Английский язык | | Русский язык | |
|---|---|---|---|---|
| | Коэффициент сложности | Пропущенные слова, % | Коэффициент сложности | Пропущенные слова, % |
| 65 | 216,1 | 1,10 | 413,3 | 7,60 |
| 100 | 224,5 | 0,65 | 481,0 | 5,31 |
| 200 | 232,4 | 0,31 | 586,8 | 2,65 |
| 400 | 236,8 | 0,17 | 670,9 | 1,19 |
| 500 | – | – | 689,9 | 0,93 |
| 800 | – | – | 713,8 | 0,64 |
| 1000 | – | – | 718,8 | 0,53 |

Кроме того, большинство словоформ одного и того же слова отличаются только в окончаниях, которые произносятся обычно не так четко, как начала слов. В результате большинство ошибок при автоматическом распознавании речи возникает именно в конце слов, что приводит к неточному пониманию всей фразы из-за несогласованности слов в ней. Поэтому основной проблемой автоматического распознавания речи для русского языка является сложный механизм словообразования, из-за которого резко возрастает размер распознаваемого словаря и падает точность. Для решения этой проблемы был введен дополнительный уровень представления речи — морфемный. На основе правил словообразования русского языка были разработаны методы автоматической обработки текстов. За счет разделения словоформы на морфемы словарь распознаваемых лексических единиц значительно сократился, так как в процессе словообразования часто используются одни и те же морфемы.

Далее рассмотрим фонетический уровень представления речи. В русском языке насчитывается 42 фонемы, из них 6 гласных и 36 согласных. В потоке речи фонемы в зависимости от их окружения могут изменять свои артикуляторно-акустические характеристики, что приводит к появлению их модификаций, называемых *аллофонами*, или оттенками фонем. Позиционные аллофоны определяются позицией данной фонемы относительно полноударного гласного. Появление комбинаторных аллофонов фонемы $\Phi_i$ связано с ее ближайшим окружением, т. е. предшествующей в потоке речи фонемой (левым контекстом) $\Phi_{i-1}$, а также последующей в потоке речи фонемой (правым контекстом) $\Phi_{i+1}$. Аллофоны фонем обозначаются тремя индексами, указывающими группу позиционных аллофонов (первый индекс), группу левого (второй индекс) и правого (третий индекс) контекстов.

При обработке речи выделяются следующие позиционные аллофоны гласных: полноударный (0), частично ударный (1), первый предударный (2), не первый предударный (3), заударный (4). Здесь в скобках указан первый индекс аллофона. С учетом левого контекста генерируются следующие комбинаторные аллофоны гласных: после синтагматической паузы (0), после переднеязычных (1), губных (2) и заднеязычных (3) твердых, после /Л/ (4), после /Р/ (5), после большинства мягких (6), после /R'/ (7), после /М'/ (8), после /Н'/ (9), после гласных (У), (О), (А), (Э), (Ы), (И). Всего — 16 левых контекстов. Здесь в скобках указан второй индекс аллофона. С учетом правого контекста генерируются следующие комбинаторные аллофоны гласных: перед синтагматической паузой (0); перед переднеязычными и заднеязычными (1); перед губными твердыми (2); мягкими (4).

Здесь в скобках указан третий индекс аллофона. Итого, при этих условиях обеспечивается генерация $N_v = N_p \cdot N_1 \cdot N_r \cdot N_{v\,ph} = 5 \cdot 16 \cdot 5 \cdot 6 = 2400$ гласных аллофонов, где $N_p$ — количество групп позиционных аллофонов; $N_1$ — количество групп левых контекстов; $N_r$ — количество групп правых контекстов; $N_{v\,ph}$ — количество гласных фонем. С учетом известных закономерностей число аллофонов гласных можно сократить до 1700 без заметного ухудшения качества синтеза речи.

В слове «*G'EN'ER'IRUJ'ECA*» («*ГЕНЕРИРУЕТСЯ*»), например, индексы ударного *I* определяются следующим образом: первый индекс (позиционный) равен 0, так как гласный полноударный, второй индекс (группа левого контекста) равен 7, так как предшествующая фонема — *R'*, третий индекс (группа правого контекста) равен 1, так как последующая фонема — *R* — принадлежит группе переднеязычных и заднеязычных твердых. Аллофоны всех гласных в слове «*G'EN'ER'IRUJ'ECA*» представлены следующим образом: «$G'E_{364}N'E_{294}R'I_{071}RU_{414}J'E_{461}CA_{410}$».

Аллофоны согласных генерируются только с учетом левого (первый индекс) и правого (второй индекс) контекстов. Левый контекст: после паузы (0), после глухих (1) и звонких (2) согласных, после гласных (3). Здесь в скобках указан первый индекс аллофона. Правый контекст: перед паузой (0), перед глухими (1) и звонкими (2) согласными, перед безударными (3) и ударными (4) гласными. Здесь в скобках указан второй индекс аллофона. Итого, при этих условиях обеспечивается генерация $N_c = N_1 \cdot N_r \cdot N_{c\,ph} = 4 \cdot 5 \cdot 36 = 720$ согласных аллофонов, где $N_{c\,ph}$ — количество согласных фонем. На практике при синтезе речи используется порядка 500 аллофонов согласных.

Например, в слове «*G'EN'ER'IRUJ'ECA*» индексы фонемы *G'* определяются следующим образом: первый индекс (группа левого контекста) равен 0, так как фонеме *G'* предшествует пауза, второй индекс (группа правого контекста) равен 3, так как последующая фонема — *E* — безударная гласная. Аллофоны всех согласных в слове «*G'EN'ER'IRUJ'ECA*» представлены следующим образом: «$G'_{03}EN'_{33}ER'_{34}IR_{33}UJ'_{33}EC_{33}A$».

## Общая структура системы фонетико-морфологической разметки речевых корпусов

Основная идея автоматизации процессов сегментации и аллофонно-морфемной маркировки заключается в реализации алгоритмов переноса меток начала и конца аллофонов с синтезированного сигнала на естественный речевой сигнал, взятый из речевых корпусов для распознавания и синтеза русской речи. Алгоритм переноса меток с одного сигнала на другой реализуется путем непрерывного во времени сопоставления естественного и синтезированного речевых сигналов с использованием алгоритмов динамического программирования (НДП-метод) [1]. Для автоматического переноса меток выбирается один из синтезированных голосов, наиболее близкий к размечаемому естественному голосу.

Общая структурная схема автоматизированной системы сегментации и аллофонно-морфемной маркировки представлена на рис. 1. На вход системы подается речевой корпус — запись естественного речевого сигнала — и орфографический текст стенограммы записи. Текст автоматически размечается на морфемы, транскрибируется и поступает на вход синтезатора речи, который, используя уже существующую БД элементов речи некоторого диктора, генерирует речевой сигнал, размеченный на фонетико-морфологические сегменты. Синтезированный и естественный речевые сигналы поступают на вход модуля переноса меток, где производится автоматическая разметка входного сигнала на фонемы и морфемы.

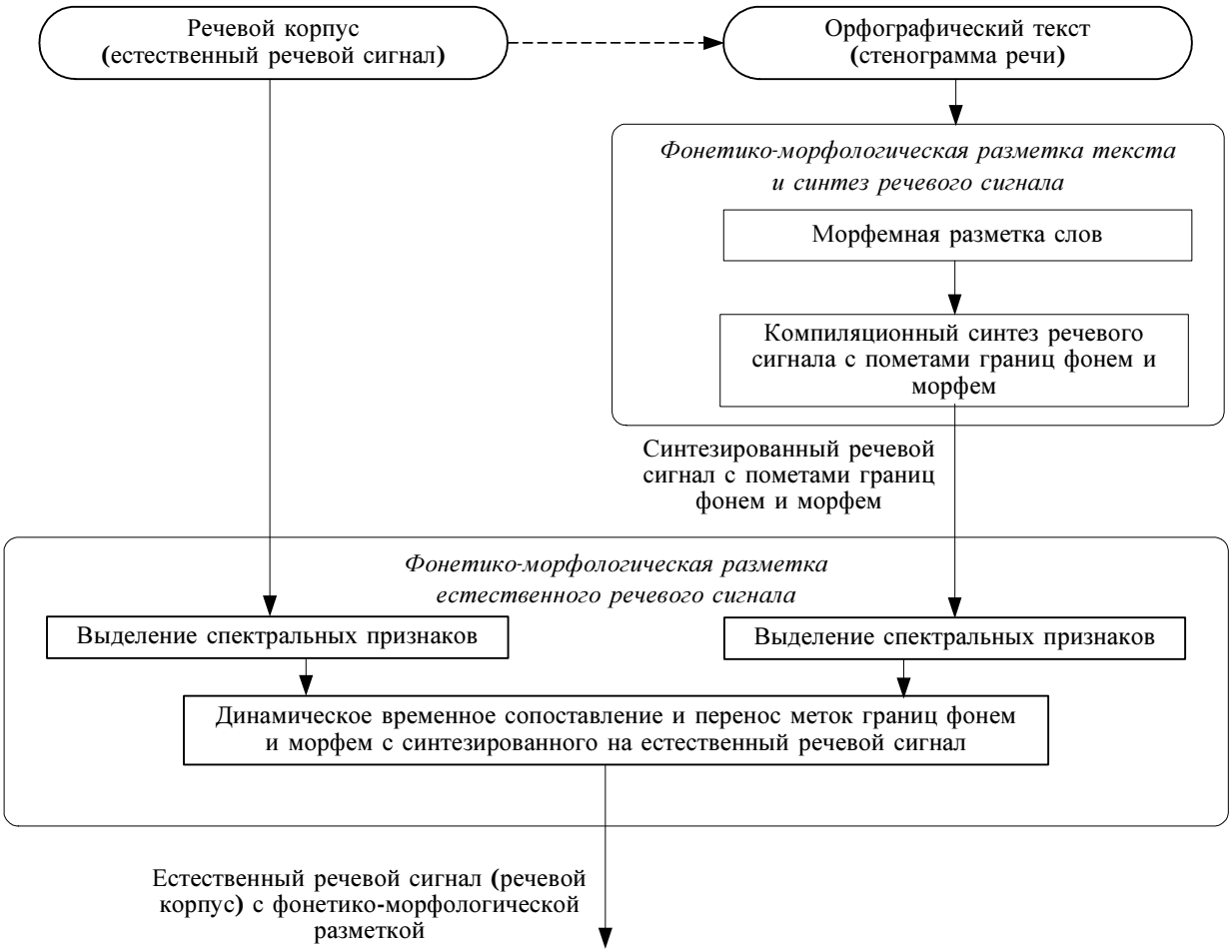
При фонетико-морфологической разметке производится автоматическая обработка текста на основе программных модулей транскрибирования и морфологического разбора. Для морфемной сегментации используются БД различных типов морфем, собранные из печатных словарей. Большая часть корневых морфем (около 4000) была взята из словаря морфем русского языка Кузнецовой [7], а различные фиксальные морфемы и флексии взяты из работы [8]. Кроме того, впоследствии, при создании ряда приложений, словарь морфем постоянно пополнялся и сейчас насчитывает около 5000 морфем.

Разбиение слова на морфемы осуществляется путем подбора различных типов морфем с учетом правил их следования в одном слове (табл. 2). Возможные пары типов морфем отмечены знаком «+». При получении недопустимой пары (отмечена в таблице знаком «–») данная гипотеза разбиения слова на морфемы откидывается и поиск продолжается дальше, пока не обнаруживается конец слова «STOP».

Разработанные БД морфем использовались для создания морфемной модели языка, строящейся на основе статистики встречаемости различных пар морфем. Для первичной оценки модели языка были использованы доступные в Интернете текстовые корпуса. Текст общим объемом около 50 Мб был предварительно обработан, а все слова в нем размечены на морфемы. В результате анализа текста были получены словарь размером около 5000 морфем и вероятности встречаемости всех пар морфем.

В результате описанного процесса автоматической обработки речевых данных получаем аннотированные корпуса речи и текста. На примере компиляционного синтезатора (рис. 2) рассмотрим их использование.

Синтезатор состоит из четырех процессоров: лингвистического, просодического, фонетического и акустического. Каждый из процессоров использует для осуществляемых им преобразований специализированные БД. В них заложены как об-



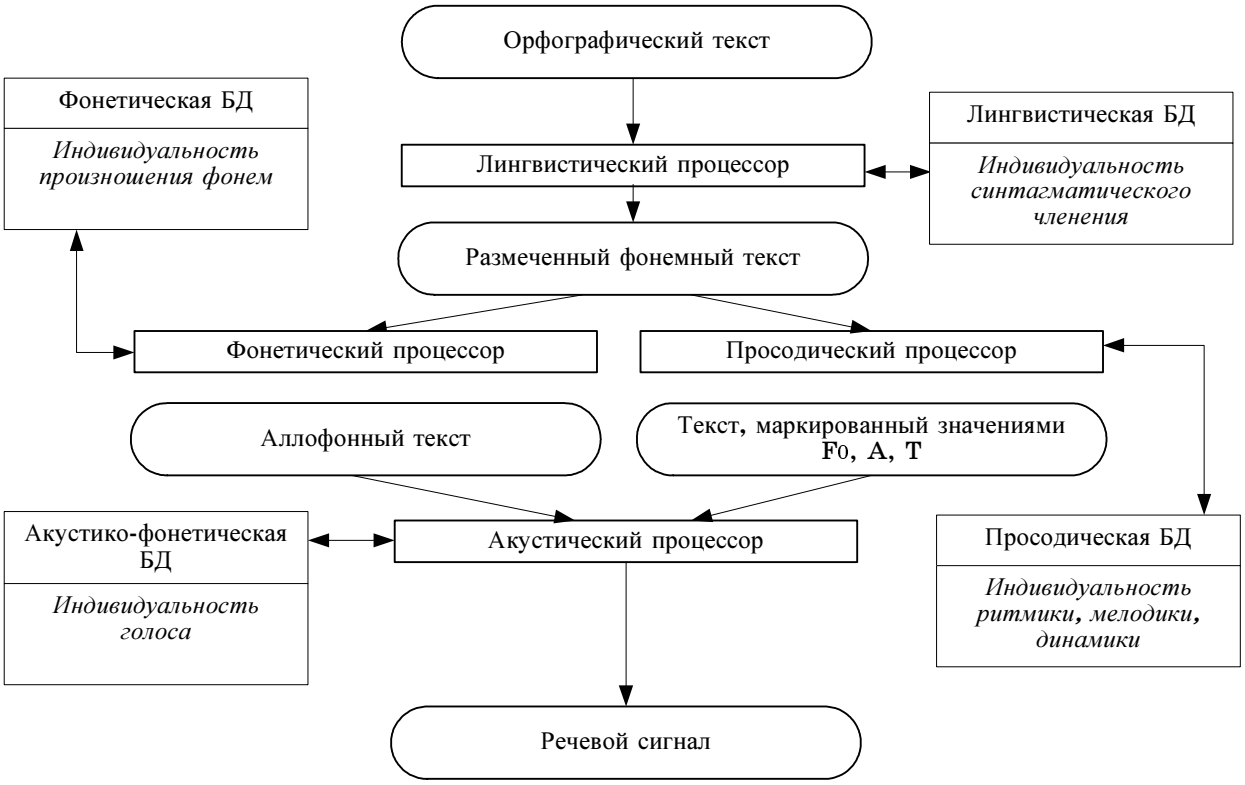

■ ***Рис. 1.*** *Общая схема системы фонетико-морфологической разметки речевых корпусов*

щие языковые правила (лингвистические, просодические, фонетические, акустические), так и связанные с индивидуальными особенностями голоса и речи диктора.

Лингвистический процессор предназначен для преобразования входного орфографического текста в размеченный фонемный текст. Под разметкой понимается разбиение текста на отдельные элементы в следующей иерархии: фонетический период, фраза, синтагма. Кроме того, процессор осуществляет расстановку словесных ударений и интонационную маркировку синтагм. Для расстановки ударений использовалась БД словоформ русского языка с информацией об ударениях в сло-

■ ***Таблица 2.*** *Таблица согласования различных типов морфем в слове*

| Тип текущей морфемы | Тип следующей морфемы | | | | |
|---|---|---|---|---|---|
| | Префикс | Корень | Интерфикс | Суффикс | Окончание |
| Нет | + | + | – | – | – |
| Префикс | + | + | – | – | – |
| Корень | STOP | STOP | + | + | + |
| Интерфикс | + | + | – | – | – |
| Суффикс | – | – | – | + | + |
| Окончание | STOP | STOP | STOP | STOP | – |




■ ***Рис. 2.*** *Общая структурная схема компиляционного синтезатора речи по тексту*

воформах, насчитывающая свыше 1 млн 700 тыс. словоформ русского языка [9].

Размеченный фонемный текст поступает на вход двух процессоров — просодического и фонетического. В результате работы просодического процессора фонемный текст делится на акцентные единицы (АЕ). Далее осуществляется разметка АЕ на элементы акцентных единиц (ЭАЕ): интонационное предъядро, ядро и заядро. И, наконец, последняя функция просодического процессора — установка значений амплитуды, длительности фонем и частоты основного тона для каждого ЭАЕ. Задача фонетического процессора заключается в генерации позиционных и комбинаторных аллофонов по входному фонемному тексту. Акустический процессор на основе информации о том, какие аллофоны требуется синтезировать и какими просодическими характеристиками должен обладать каждый аллофон, генерирует синтетический речевой сигнал. При этом используется БД, в которой хранятся акустические волны аллофонов синтезируемого голоса. Полученный синтезированный сигнал используется для фонетико-морфологической разметки естественного речевого сигнала.

Модуль разметки включает в себя анализатор спектральных параметров естественного и синтезированного сигналов. Затем на основе выделенных параметров выполняется динамическое временное сопоставление естественного и синтезированного речевых сигналов. По найденному соответствию производится перенос меток с синтезированного на естественный речевой сигнал.

Анализ параметров речевого сигнала в описываемой системе заключается в вычислении нормированной сонограммы. Вычисление сонограммы проводится путем пропускания исходной записи через набор полосовых фильтров Чебышева и вычисления среднего значения сигналов на каждом участке длительностью 10 мс. Полосы пропускания фильтров выбраны в соответствии со шкалой Барка. Нормирование сонограммы осуществляется по формуле

$$Sn(n, j) = \sum_{k=n-T}^{n+T} \sum_{l=0}^{C} \Delta(S(n, j), S(k, l)),$$

где $Sn(n, j)$, $S(n, j)$ — нормированное и ненормированное значения точки $n, j$ сонограммы соответственно; $T$ — интервал нормирования; $C$ — число каналов в сонограмме.

Функция

$$\Delta(S(n, j), S(k, l)) = \begin{cases} 1, & \text{если } S(n,j) - S(k,l) > \varepsilon \\ 0, & \text{если } -\varepsilon \le S(n, j) - S(k, l) \le \varepsilon, \\ -1, & \text{если } S(n, j) - S(k, l) < -\varepsilon \end{cases}$$

где ε — порог, определяемый уровнем шума во входном сигнале.

После вычисления сонограмм синтезированного и естественного речевых сигналов их сравнива-

ют НДП-методом, в основу которого положен итерационный алгоритм вычисления интегральных расстояний:

$$D(n,m)=\min\begin{bmatrix} D(n-1,m)+k_h d\{\overline{S(n)};\overline{E(m)}\}+ \\ D(n,m-1)+k_v d\{\overline{S(n)};\overline{E(m)}\}+ \\ D(n-1,m-1)+k_d d\{\overline{S(n)};\overline{E(m)}\}+ \\ +\frac{k}{M}|m-T(n-1,m)|; \\ +\frac{k}{M}|m-1-T(n,m-1)|; \\ +\frac{k}{M}|m-1-T(n-1,m-1)|. \end{bmatrix}$$

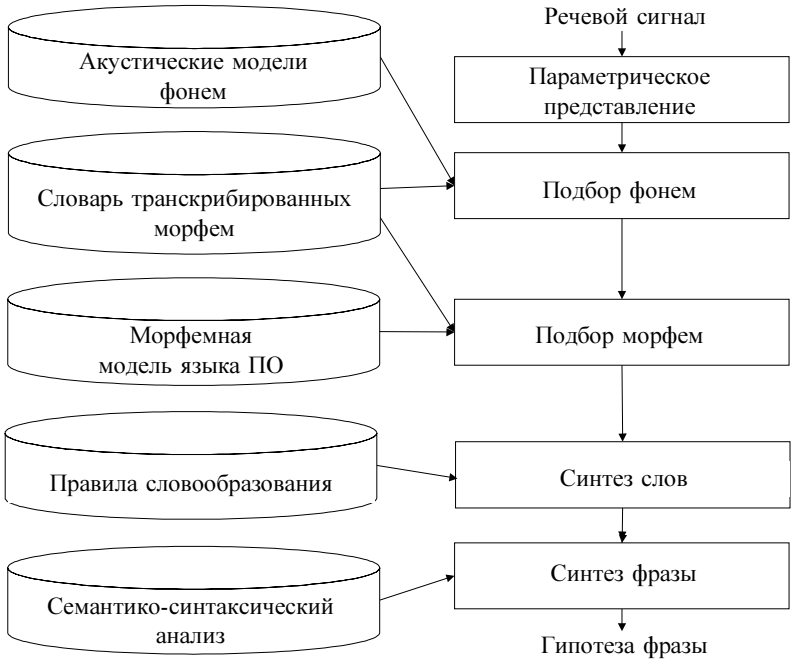
При использовании классических формул НДП результаты сравнения сигналов были не всегда точными, поэтому в описываемой системе в эти формулы включены дополнительные коэффициенты: коэффициент времени $k$ и множители $k_h, k_v, k_d$, причем коэффициент $k$ нормируется по отношению к длине синтезированного сигнала $M$. Оптимальные значения коэффициентов $k_h, k_v, k_d, k$ были найдены экспериментальным путем.

## Применение фонетико-морфологической разметки для распознавания речи

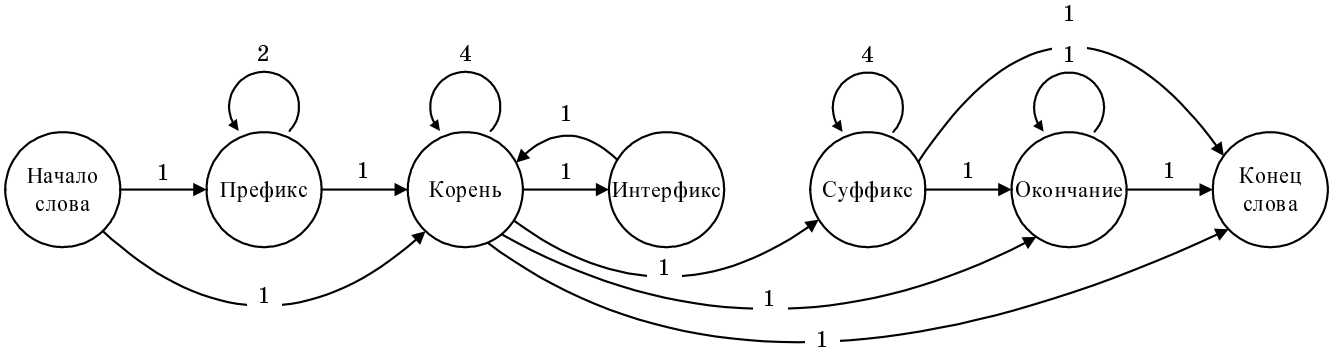
Тестирование системы распознавания и полученных БД производится в задачах голосового доступа к рубрикам электронного каталога «Желтые страницы Санкт-Петербурга» (размер словаря составил 1850 слов), а также заказа авиабилетов из аэропортов Санкт-Петербурга. Далее приводятся результаты распознавания именно по первой задаче — распознавания названий рубрик. В общепринятую архитектуру распознавания речи был введен дополнительный уровень представления языка и речи — морфемный и разработана оригинальная система распознавания русской речи SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) [5] (рис. 3).

Речевой сигнал, поступающий с микрофона, в первую очередь проходит этап параметрического представления, где отрезаются начальные и конечные паузы в сигнале, а оставшийся участок кодируется в последовательность векторов признаков, которая уже следует в модуль распознавания фонем. При распознавании фонем (которые используются в форме трифонов) и формировании морфем используются методы скрытого марковского моделирования и смесей гауссовских распределений. В отличие от существующих аналогов, в нашей модели вместо слов используются морфемы. За счет этого на этапе распознавания лексических единиц было получено существенное увеличение скорости.

После распознавания фонем и подбора наиболее вероятных цепочек морфем получившийся набор гипотез далее используется для формирования цепочек слов. Синтез слов из различных типов морфем осуществляется по схеме, представленной на рис. 4. В данной модели заданы начальное и конечное состояния, а в остальных узлах присутствуют различные типы морфем. Дугами обозначены возможные переходы. Планируется сделать эту




■ ***Рис. 3.*** *Структура системы распознавания русской речи SIRIUS*




■ ***Рис. 4.*** *Синтез слова из морфем*

■ ***Таблица 3.*** *Сравнение моделей распознавания по точности*

| Диктор | Точность целословного распознавания | | Точность морфемно-ориентированного распознавания | |
|---|---|---|---|---|
| | слов | фраз | морфем | фраз |
| 1 | 96 | 94 | 82 | 92 |
| 2 | 93 | 90 | 78 | 90 |
| 3 | 95 | 93 | 81 | 92 |
| 4 | 91 | 91 | 79 | 91 |
| 5 | 95 | 93 | 82 | 91 |
| Среднее | 94,0 | 92,2 | 80,4 | 91,2 |

модель вероятностной, а пока максимальное количество переходов из состояния в состояние заданы жестко. На этом этапе обработки на основе каждой поступившей гипотезы фразы, представленной в виде последовательности морфем, формируется еще несколько гипотез, представленных последовательностью гипотез слов.

Последним этапом обработки является синтез фразы. На входе этого уровня мы получаем цепочку слов, составляющую произнесенное высказывание. Однако процент ошибок распознавания слов здесь достаточно высок, и большинство из них, как уже отмечалось, происходит из-за ошибок в распознавании окончаний, которые произносятся не так четко, как начала слов. А ошибки в окончаниях при распознавании слов приводят к тому, что происходит ошибка в распознавании всей фразы из-за несогласованности слов в предложении. Для того чтобы исправить эти ошибки, мы используем морфологический анализ предложения и грамматические правила русского языка. Таким образом, на выходе системы распознавания речи мы получаем цепочку слов, составляющих предложение.

Для тестирования системы использовались 635 фраз, записанных в офисных условиях. В эксперименте участвовало 5 дикторов. Записанные файлы были пропущены через модель целословного распознавания, а затем через морфемно-ориентированную модель распознавания с последующим словообразованием (табл. 3). По сравнению с первой моделью точность распознавания морфем несколько снизилась, но за счет последующих уровней обработки точность распознавания фраз практически не изменилась.

■ ***Таблица 4.*** *Сравнение моделей распознавания по скорости обработки*

| Распознавание | Время, затрачиваемое на тестовый набор, с | Среднее время, с | | |
|---|---|---|---|---|
| | | на одну фразу | на одно слово | на одну морфему |
| Целословное | 2993 | 4,71 | 1,16 | – |
| Морфемно-ориентированное | 1740 | 2,74 | 0,67 | 0,47 |

Также был проведен тест по сравнению скорости работы морфемного и целословного распознавателей (табл. 4). Общее количество тестовых фраз, содержащихся в тестовой БД, составило 635 (состоящих из 2574 слов). Из таблицы видно, что при использовании разработанной системы скорость возросла более чем в 1,7 раза, что при незначительном падении точности позволяет говорить о создании перспективной системы распознавания речи для больших словарей.

Таким образом, разработанный модуль распознавания слитной русской речи показал достаточно высокую точность и дикторонезависимость к носителям русского языка.

## Применение фонетико-аллофонной разметки для персонализированного синтеза речи по тексту

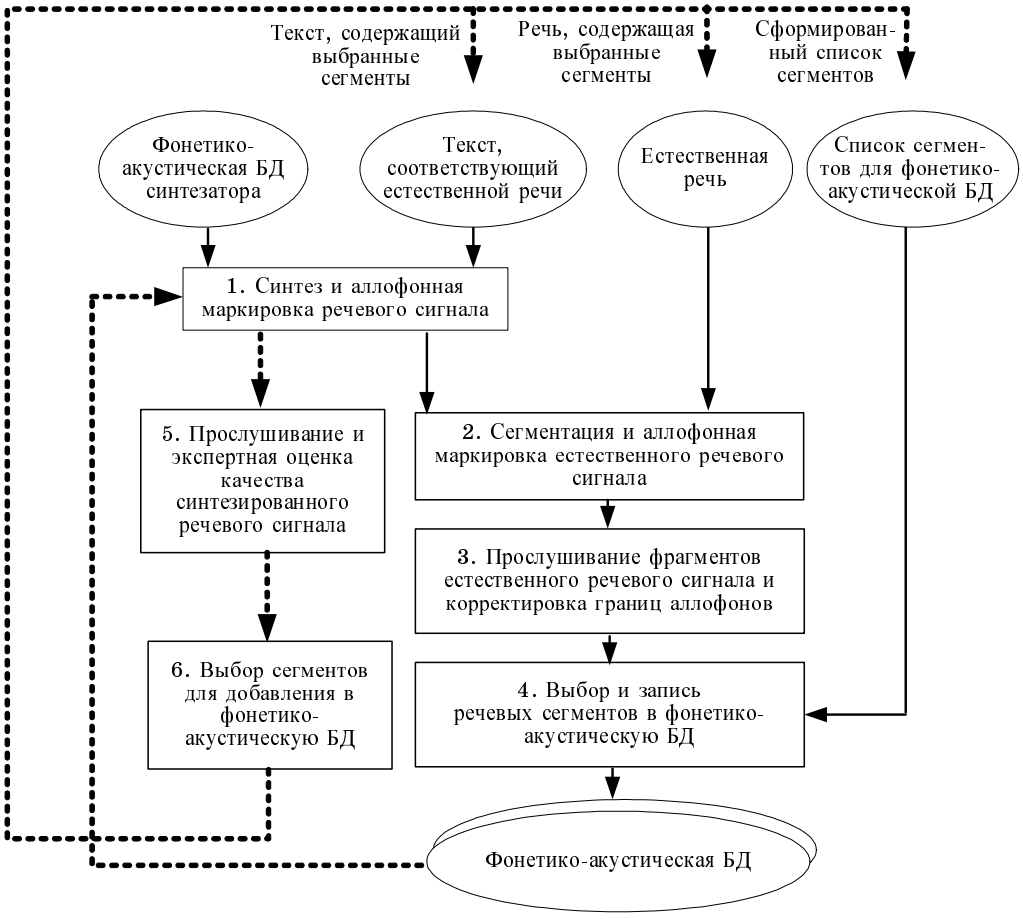
Для автоматизированного создания БД фонетико-акустических характеристик голоса и дикции личности создана система AcousticClonator.

В системе реализованы этапы создания индивидуализированных БД, ответственных за синтез акустических свойств голоса и фонетических особенностей произношения.

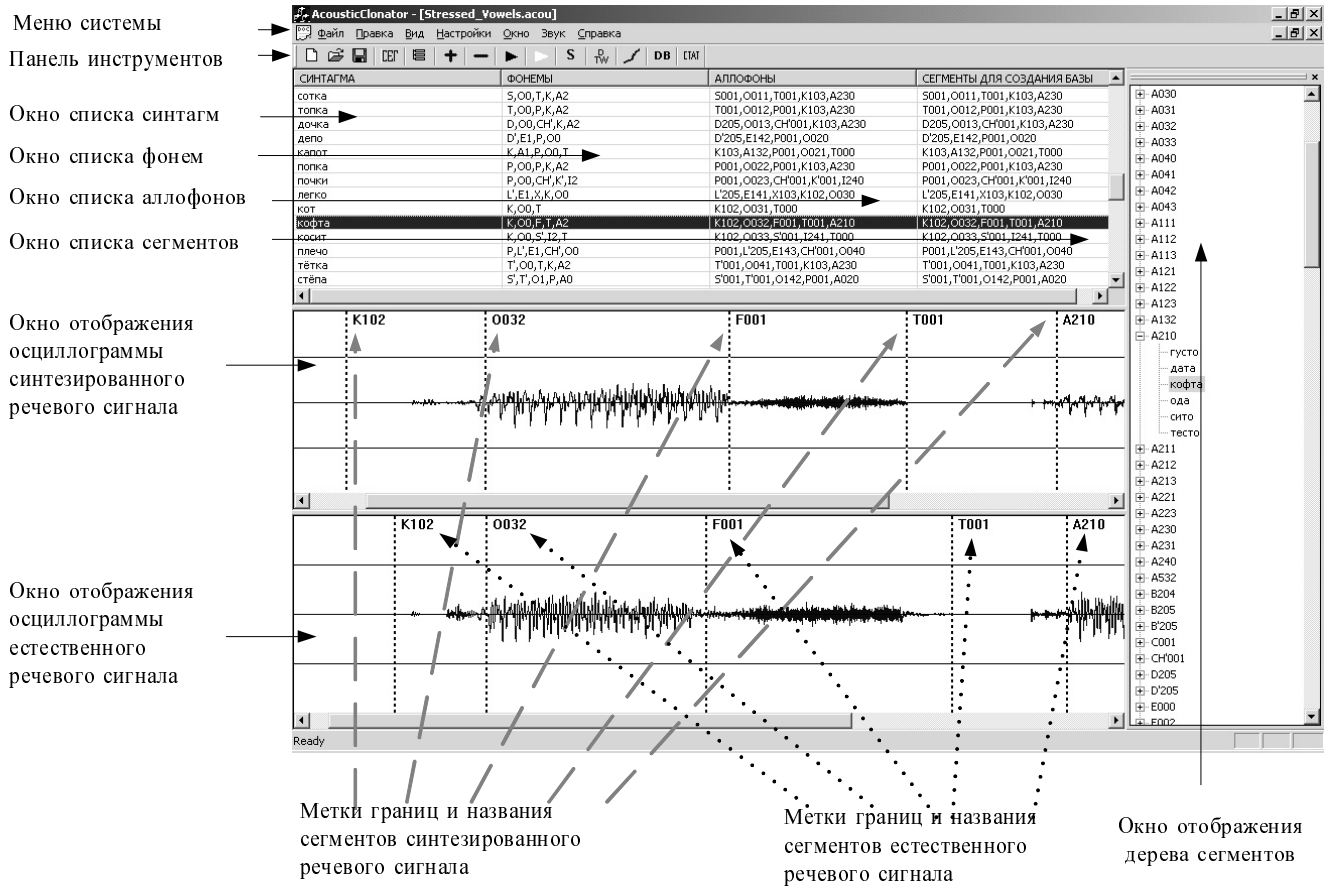
Технология создания фонетико-акустической БД с использованием этой системы, взаимодействие блоков, входные и выходные данные показаны на рис. 5. Создание фонетико-акустической БД состоит из двух или более этапов. Последовательность действий, выполняемых на первом этапе, обозначена сплошными линиями, а на втором и последующих этапах — штриховыми.

На первом этапе в качестве списка сегментов для фонетико-акустической БД используется минимально необходимый или расширенный набор аллофонов, а в качестве речевой базы — звуковой массив, включающий набор русских слов, содержащий все необходимые аллофоны. Текст, соответствующий звуковому массиву, а также существующая фонетико-акустическая БД являются входными данными для синтеза и аллофонной маркировки речевого сигнала *1*. Аллофонно-размеченный синтезированный речевой сигнал используется для сегментации и аллофонной маркировки естественного речевого сигнала *2*. Следующая функция *3* — прослушивание и, при необходимости, ручная корректировка границ аллофонов — реализуется экспертом-фонетистом, который может посмотреть осциллограмму речевого сигнала, прослушать любой его участок, передвинуть метки границ аллофонов. И, наконец, пользователь может прослушать все слова, содержащие указанный речевой сегмент, и установить, из какого именно слова указанный сегмент будет помещен в БД *4*. Результатом первого этапа является фонетико-акустическая БД нового диктора, содержащая минимально необходимый (или расширенный) набор звуковых волн аллофонов.

На втором этапе в синтезаторе речи используется вновь созданная фонетико-акустическая БД. Эксперт-фонетист прослушивает синтезированный по различным текстам речевой сигнал, оценивает его качество, выделяет речевые участки, качество звучания которых неудовлетворительно, и формирует список сегментов для пополнения фонетико-акустической БД. В качестве таких сегментов могут выступать аллофоны, мультифоны (сочетания аллофонов), слоги и другие сегменты речи. Сформированный список сегментов для пополнения фонетико-акустической БД, дополни-



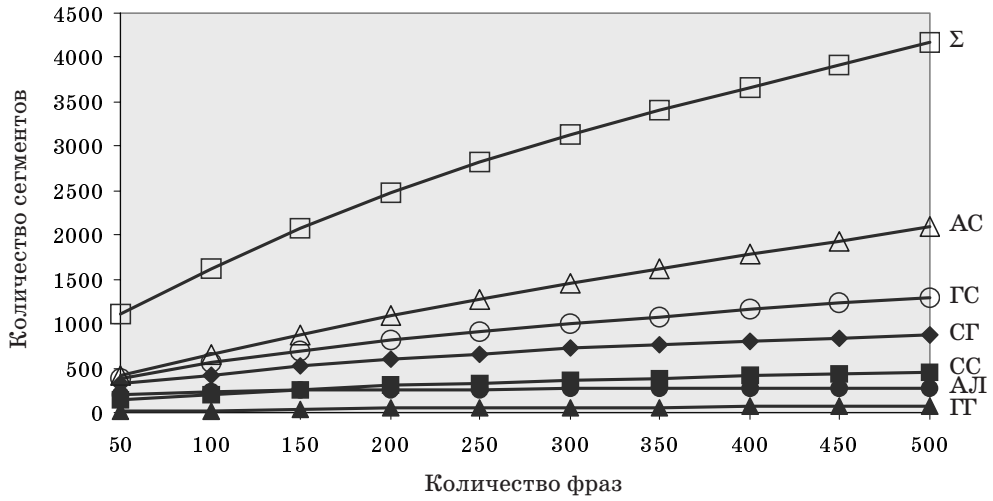

■ ***Рис. 5.*** *Технология создания фонетико-акустической БД*




■ ***Рис. 6.*** *Общий вид и основные блоки пользовательского интерфейса системы*

тельная речевая база, содержащая эти сегменты, и соответствующий текст снова подаются на вход системы. В качестве дополнительных речевых баз используются записи двух специально подобранных фонетически репрезентативных текстов.

Второй этап создания фонетико-акустической БД может повторяться несколько раз, при этом после очередного прослушивания синтезированного речевого сигнала и оценки его качества эксперт может формировать все новые списки сегментов




■ ***Рис. 7.*** *Графики зависимости количества аллофонов (АЛ), диаллофонов (ГГ, СГ, СС, ГС), аллослогов (АС) и общего количества сегментов (Σ) от объема выборки*

для пополнения БД, с каждым шагом повышая качество синтезируемой речи.

Пользовательский интерфейс системы (рис. 6) реализован в среде MS Visual C++ 6.0. Приложение имеет мультидокументный тип, что дает возможность пользователю варьировать наборы входных данных и делает систему более гибкой и легко настраиваемой. Основные блоки пользовательского интерфейса состоят из меню, панели инструментов, блока окон, в котором отображается список синтагм (орфографический текст), осциллограммы естественного и синтезированного речевых сигналов, синтагматического дерева сегментов.

Описанная система сегментации речевого сигнала используется для создания БД элементов речи, которая необходима для распознавания, компьютерного клонирования голоса и речи нового диктора, а также для наращивания уже существующих БД дикторов путем добавления в базу новых элементов синтеза — диаллофонов и аллослогов, что позволяет значительно улучшить качество синтезируемой речи. С использованием разработанной системы была создана акустическая БД речевых сегментов различной размерности — аллофонов, диаллофонов и аллослогов. В качестве текстовой БД использованы таблицы ГОСТ для измерения фразовой разборчивости речи [10], разработанные с учетом фонетической репрезентативности русской речи. Таблицы включают 500 фонетически сбалансированных фраз. На рис. 7 представлены графики зависимости количества различных аллофонов; диаллофонов типа ГГ (гласный-гласный), СГ (согласный-гласный), СС (согласный-согласный) и ГС (гласный-согласный), а также аллослогов, которое можно получить с использованием описанной системы в зависимости от объема выборки фраз из текстовой БД ГОСТ.

## Заключение

Исследование проблем автоматической обработки речи является важным фундаментальным направлением. Эти проблемы сдерживают развитие всевозможных систем взаимодействия человека с машиной. Представленные модели синтеза и распознавания речи в первую очередь направлены на учет особенностей русского языка. Авторами разработан и опробован новый морфемный метод представления языка и речи. Он показал высокое качество и устойчивость работы на словаре до 2000 слов конкретной предметной области. Разработаны БД различных типов морфем. В результате такой обработки обеспечивается инвариантность к грамматическим отклонениям, увеличивается скорость распознавания русской речи и других языков со сложным механизмом словообразования (в частности, славянских). Последующие работы направлены на увеличение размера распознаваемого словаря, настройку системы распознавания русской речи SIRIUS к работе с телефонным каналом. При внедрении системы распознавания речи в телекоммуникационные приложения будут учтены проблемы, связанные со спецификой телефонных линий и различиями характеристик телефонных аппаратов. Накопленные речевые и лексические БД будут использованы для дальнейшего изучения механизма понимания речи и создания эффективных средств человеко-машинного взаимодействия.

Разработанные системы сегментации речевого сигнала и текста позволяют автоматизировать трудоемкий процесс создания акустических БД, а также учитывать индивидуальные фонетико-акустические особенности голоса личности. Как показал опыт, создание минимально необходимого набора аллофонов опытным фонетистом вручную занимает несколько недель рабочего времени, а расширенного набора — более двух месяцев. При использовании предложенной системы создание БД, содержащей полный набор звуковых волн аллофонов, занимает не более одного часа машинного времени. Автоматизация процесса пополнения БД мультифонами (диаллофонами и аллослогами) позволила значительно повысить качество синтезированной речи, практически приблизив к естественной. Сказанное подтверждается результатами испытания новой системы синтеза речи по тесту «МУЛЬТИФОН», акустическая БД которой содержит свыше 4 тыс. фонетических сегментов речи, полученных с помощью описанной системы фонетико-аллофонной разметки речевого сигнала. Кроме очевидного практического применения – создания индивидуализированных речевых БД для последующего высококачественного синтеза речи по тексту с манерой чтения конкретного человека и его голосом, система клонирования фонетико-акустических характеристик речи может использоваться в криминалистике для экспресс-идентификации голоса подозреваемого [11].

Данные исследования проводятся при финансовой поддержке правительства Санкт-Петербурга, Европейского Сообщества SIMILAR NoE FP6: IST-2002-507609, а также проекта INTAS № 04-77-7404.

## Литература

1. **Давыдов А., Киселев В., Лобанов Б., Цирульник Л.** Система сегментации речевого сигнала методом анализа через синтез // Изв. Белорусской инженерной академии. 2004. № 1 (17)/1. С. 112–114.
2. **Лобанов Б. М.** Компьютерное «клонирование» персонального голоса и речи // Новости искусственного интеллекта. 2002. № 5 (55). С. 35–39.
3. **Lobanov B. M., Tsirulnik L. I.** Phonetic-Acoustical Problems of Personal Voice Cloning by TTS: Proc. of the International Conference SPECOM'2004. St. Petersburg, 2004. P. 17–21.

4. **Lobanov B. M., Tsirulnik L. I.** Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian): Proc. of the International Conference SPECOM'2006. St. Petersburg, 2006. P. 274–283.
5. **Карпов А. А., Ронжин А. Л., Ли И. В.** SIRIUS — система дикторонезависимого распознавания слитной русской речи // Изв. ТРТУ. 2005. № 10. С. 44–53.
6. **Whittaker E. W. D.** Statistical Language Modelling for Automatic Speech Recognition of Russian and English. PhD thesis / Cambridge University. Cambridge, 2000. P. 141.
7. **Кузнецова А. И., Ефремова Т. Ф.** Словарь морфем русского языка: Ок. 52000 слов. М.: Рус. яз., 1986. 1136 с.
8. **Русская** грамматика: В 2 т. / Редкол.: Н. Ю. Шведова (гл. ред.) и др. Т. 1: Фонетика. Фонология. Ударение. Интонации. Словообразование. Морфология / Н. С. Авилова, А. В. Бондарко, Е. А. Брызгунова и др. М.: Наука, 1980. 783 с.
9. **http:**//starling.rinet.ru
10. **ГОСТ 16600–72.** Передача речи по трактам радиотелефонной связи. М.: Изд-во стандартов, 1973.
11. **Давыдов А. Г., Киселев В. В., Лобанов Б. М., Цирульник Л. И.** Система экспресс-идентификации голоса личности методом клонирования акустических характеристик речи // Теория и практика речевой коммуникации: Тез. докл. Междунар. конф. М., 2004. С. 23–28.

## «IV МЕЖВУЗОВСКАЯ КОНФЕРЕНЦИЯ МОЛОДЫХ УЧЕНЫХ» апрель 2007 г.

**Место проведения конференции: Санкт-Петербургский государственный университет информационных технологий, механики и оптики.**
**Адрес: 197101, Санкт-Петербург, Кронверкский пр., д. 49**

**Цель конференции**

Конференция проводится с целью стимулирования научно-технической деятельности молодых ученых, приобретения ими опыта публичных выступлений, подачи научных документов для публикации, а также с целью ознакомления научной общественности с результатами исследований приоритетных направлений развития науки, технологий и техники.

К участию приглашаются молодые ученые (до 35 лет).

**Программа конференции**

Планируется проведение пленарного заседания и секций с устными докладами.

В рамках конференции будут проведены научные школы: I сессия научной школы «Информационные технологии в образовании», II сессия научной школы «Информационная безопасность, проектирование, технология элементов и узлов компьютерных систем».

**Направления работы конференции**

Информационные технологии
Информационно-телекоммуникационные системы
Безопасность и противодействие терроризму, защита информации
Технологии производства программного обеспечения
Технологии приборостроения, мехатроника и робототехника
Системный анализ, математическое моделирование и управление в технических системах
Теплофизика и теоретическая теплотехника
Живые системы, биомедицинские технологии и томография
Оптотехника и оптические материалы
Фотоника и оптоинформатика
Физика твердого тела, наносистем и материалов
Гуманитарные науки (философия, социология, политология, педагогика)
Экономика, финансы и менеджмент организации

**Контрольные сроки**

Тезисы докладов и тексты статей для сборника принимаются до **2 февраля 2007 г.**

**Издание трудов конференции**

По результатам работы конференции планируется выпуск сборника тезисов докладов и сборника лучших докладов.

**Дополнительная информация**

Оргкомитет конференции,
тел.: (812) 232-04-64,
эл. почта: kmu@mail.ifmo.ru
сайт: http://www.ifmo.ru/index.php?out=itmo_science_conference_07