На правах рукописи

Калюжный Михаил Васильевич

СИСТЕМА РЕАБИЛИТАЦИИ СЛАБОВИДЯЩИХ НА ОСНОВЕ НАСТРАИВАЕМОЙ СЕГМЕНТАРНОЙ МОДЕЛИ СИНТЕЗИРУЕМОЙ РЕЧИ

Специальность: 05.11.17 – Приборы, системы и изделия медицинского назначения

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2009

Работа выполнена в Тверском государственном техническом университете.

Научный руководитель –
доктор технических наук, профессор Филатова Наталья Николаевна.

Официальные оппоненты:
доктор технических наук, профессор Аббакумов Константин Евгеньевич;
доктор технических наук, доцент Вахитов Шакир Яшэрович.

Ведущая организация – Научно-производственное предприятие "Межотраслевой центр эргономических исследований и разработок" (НПП "ЭРГОЦЕНТР"), г. Тверь

Защита диссертации состоится “13” мая 2009 г. в 11 часов на заседании совета по защите докторских и кандидатских диссертаций Д 212.238.09 Санкт-Петербургского государственного электротехнического университета “ЛЭТИ” имени В. И. Ульянова (Ленина) по адресу: 197376, Санкт-Петербург, ул. Проф. Попова, 5, ауд. 5652.

С диссертацией можно ознакомиться в библиотеке университета.

Автореферат разослан “02” апреля 2009 г.

Учёный секретарь совета
по защите докторских
и кандидатских диссертаций Болсунов К.Н.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Современное общество, следуя в своём развитии принципам гуманизма, должно обеспечивать равные права и возможности каждому человеку. Поэтому актуальной проблемой является реабилитация инвалидов и просто людей с ослабленным здоровьем.

Развитие информационных технологий и распространение персональных компьютеров привело к их повсеместному использованию в качестве рабочего инструмента и домашнего многофункционального бытового прибора. Это, с одной стороны, позволяет людям с ограниченными физическими возможностями более полноценно трудиться и жить более активной жизнью, а с другой стороны, делает актуальной проблему использования компьютера такими людьми. Решение данной проблемы видится в создании специальных технических средств реабилитации (ТСР), позволяющих инвалидам эффективно и комфортно работать с компьютером.

Основной задачей реабилитации слепых и слабовидящих является восстановление информационного обмена между личностью и внешним миром, нарушенного вследствие зрительной патологии.

Анализ современных систем реабилитации позволяет выделить два основных подхода: коррекционный и компенсационный.

При реабилитации людей с нарушением зрения первый подход реализуется с помощью ТСР, позволяющих увеличить резкость, размер или контрастность изображения и тем самым дать возможность человеку воспринимать визуальную информацию. К таким средствам относятся очки, контактные линзы, оптические увеличители и т.п. Второй подход, подразумевающий использование других каналов восприятия – осязания и слуха, построен на применении специальных рельефных изображений и шрифтов, а также звуковых сигналов, главным образом, речи.

Применение синтеза речи в средствах реабилитации незрячих и слабовидящих, сдерживается недостаточным качеством получаемого речевого сигнала (РС). Основными показателями качества синтезированной речи являются естественность и разборчивость. РС современных синтезаторов, обладая хорошей разборчивостью, звучит недостаточно естественно. Это приводит к быстрому утомлению пользователя и понижает эффективность ТСР.

Причина низкой естественности заключается в несоответствии ритмо-интонационных характеристик синтезируемой речи соответствующим характеристикам естественной речи.

Проблеме синтеза естественно звучащей речи посвящены работы И. Алдошиной, А.В. Бабкина, В.И. Галунова, О.Ф. Кривновой, Б.М. Лобанова, Р.К. Потаповой, F. Charpentier, Sh. Naraynan, где отмечено влияние эмоциональных проявлений в речи на естественность её звучания и отсутствие решений, позволяющих синтезировать эмоционально окрашенную речь.

Приведенные факты позволяют сделать вывод об актуальности темы диссертации, которая определяется:

- отсутствием эргономичных речевых интерфейсов, позволяющих незрячим и слабовидящим людям эффективно и комфортно пользоваться компьютером;

- отсутствием моделей и алгоритмов синтеза естественно звучащей эмоционально окрашенной речи;

- существующими тенденциями развития и потребностями общества, делающими пользование компьютером существенным условием для полноценной жизни и профессиональной деятельности;

- возможностями компьютерных технологий и современных методов обработки информации.

Цели и задачи работы.

Целью работы является создание моделей, алгоритмов и программного обеспечения, позволяющего синтезировать естественно звучащий речевой сигнал, и разработка на их основе системы реабилитации слабовидящих.

Задачи исследования. Для достижения поставленной цели в диссертации необходимо решить следующие задачи:

1. Выполнить функционально-структурный анализ существующих подходов к решению задачи реабилитации незрячих и слабовидящих, а также способов их реализации в современных ТСР.

2. Выполнить анализ роли и места средств речевого вывода в ТСР, способов формирования РС и методов оценки качества РС.

3. Определить параметры, позволяющие количественно описывать характеристики, влияющие на качество сигнала. Разработать алгоритмы оценки и модификации параметров, определяющих различие естественного и искусственного речевых сигналов.

4. Разработать методику и провести экспериментальные исследования с целью получения образцов РС с заданными характеристиками, а также с целью оценки характеристик естественных и модифицированных РС.

5. Разработать алгоритмическое и программное обеспечение для анализа и коррекции просодических характеристик РС, обеспечивающее синтез естественно звучащей эмоционально окрашенной речи.

6. Разработать архитектуру программной системы реабилитации слабовидящих на основе созданных моделей и алгоритмов синтеза естественно звучащей речи, выполнить экспериментальную проверку новых моделей и алгоритмов.

Объектом исследования является метод синтеза естественно звучащей речи в системах реабилитации незрячих и слабовидящих.

Предметом исследования является информационное, методическое, алгоритмическое и программное обеспечение для коррекции просодических характеристик, позволяющей повысить качество синтезируемой речи.

Методы исследования. Для решения поставленных задач в качестве базовой методологии, являющейся основой исследования, в работе использовались методы структурного системного анализа. Также в работе использовались методы обработки сигналов, теории вероятностей и математической ста-

тистики, теории нечётких множеств, методы кластерного анализа, теории биотехнических систем и элементы психологии эмоций.

Новые научные результаты:

1. Информационная модель просодии, описывающая взаимосвязи между факторами, характеристиками и параметрами для естественного и для синтезируемого речевого сигнала.
2. Экспериментальная методика получения образцов РС, различающихся по эмоциональному окрасу.
3. Методика пофонемной оценки эмоциональности речевого сигнала.
4. Экспериментально подтверждённая гипотеза о локализации эмоциональной компоненты на гласных и вокализованных звуках РС.
5. Сегментарная модель представления вокализованных участков РС.

Практическую ценность работы составляют:

1. Алгоритм коррекции эмоционального окраса речи путём изменения параметров сегментарной модели РС.
2. Программное обеспечение для коррекции просодических, в т.ч. эмоциональных характеристик РС на основе сегментарной модели, позволяющее проводить сегментацию РС, вычислять параметры шаблона, их приращения и отношения, редактировать значения параметров, работать с файлами параметров, синтезировать РС по заданным параметрам.
3. Компоненты ПО для речевых движков, реализующие коррекцию эмоциональных характеристик при синтезе речи.
4. Результаты экспериментальных исследований, подтверждающие достоверность предложенных методик, моделей и алгоритмов.
5. Архитектура программной системы реабилитации слабовидящих на базе приложения типа «Голосовой менеджер» и речевого движка, реализующего синтез речи на основе настраиваемой сегментарной модели.

Внедрение результатов.

Результаты диссертационной работы внедрены в ОАО НПП "ЭРГОЦЕНТР" (г. Тверь); внедрены в НПО «Вымпел» (г. Тверь); создан учебный стенд, используемый в учебном процессе Тверского государственного технического университета.

Апробация результатов работы. Научные и практические результаты диссертационной работы докладывались и обсуждались в 2005-2008 годах на V Международной научно-технической конференции «Электроника и информатика-2005» (МИЭТ, Зеленоград, 2005), на «Научной сессии МИФИ-2008» (МИФИ, Москва, 2008) и на XXI Международной НТК «Математические методы в технике и технологиях (ММТТ-21)» (СГТУ, Саратов, 2008).

Основные положения, выносимые на защиту:

1. Возможно управление эмоциональной характеристикой синтезируемой речи путём изменения параметров гласных фонем.

2. Сегментарная модель позволяет описывать гласные участки речевого сигнала без потери качества.

3. Методика коррекции параметров шаблонных сегментов позволяет изменять эмоциональный окрас речи, сохраняя индивидуальные особенности голоса.

Публикации. Основные теоретические и практические результаты диссертации опубликованы в 7 работах, среди которых 1 публикация в ведущих рецензируемых изданиях, рекомендованных в действующем перечне ВАК, а также 2 статьи в других журналах и изданиях, 3 публикации в трудах международных научно-технических конференций, Основные положения защищены 1 свидетельством на программу для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, 4 глав с выводами, заключения, списка литературы и приложений. Основное содержание работы изложено на 137 страницах машинописного текста, 32 рисунках, 29 таблицах, 2 приложениях. Список использованной литературы включает 69 наименований, среди которых 38 отечественных и 31 иностранных авторов.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи исследований, приведено краткое содержание работы по главам.

В первой главе диссертации выполнен функционально-структурный анализ технических средств реабилитации (ТСР) незрячих и слабовидящих, используемых при работе на компьютере, рассмотрена реализация компенсационного и коррекционного подходов.

Дополнительно к описанным в литературе аспектам реабилитации (медицинскому, физическому, психологическому, профессиональному и социально-экономическому) рассмотрен технический аспект – разработка и внедрение технических средств, используемых в целях реабилитации.

Предложена классификация методов, реализуемых в ТСР незрячих и слабовидящих. Выделены визуальные (увеличение размера и повышение контрастности изображения), тактильные (использование шрифта Брайля и рельефных изображений) и речевые (воспроизведение записанных речевых сообщений и синтез речи по тексту) методы. Первая группа методов отнесена к коррекционным, две других – к компенсационным методам реабилитации.

Показано, что использование речевых технологий как инструмента для реабилитации слабовидящих наиболее перспективно. Речь является универсальным способом взаимодействия программ и технических средств с незрячим или слабовидящим пользователем. Брайлевские устройства вывода (дисплеи и принтеры) уступают речевому выводу информации по скорости и доступности, так как требуют от пользователя умения читать рельефно-точечный шрифт, сложны в освоении и недешевы.

Выделено 3 класса ТСР: аппаратные, программные и автономные.

Среди существующих программных ТСР перспективными являются 2 типа приложений: программы экранного доступа (ПЭД) и голосовые менеджеры (ГМ). Проведен сравнительный анализ ряда ГМ, в том числе: «JAWS»; «VIRGO»; «Говорящая мышь (Speaking Mouse Home)» v. 4.6; «VoiceMan» (+L&H TTS Engine Russian); «Talker» (+Sakrament TTS Engine Russian v.2.5); «DigIt Clipboard Reader» по следующим критериям: функциональность, свойства голосового модуля, свойства интерфейса, дополнительные возможности.

Анализ показал, что для программных ТСР наиболее эффективным является использование внешних голосовых модулей, устанавливаемых как компонент операционной системы (ОС), или речевых движков (TTS-engine). Данный подход имеет ряд преимуществ, в том числе: избавляет разработчиков ТСР от необходимости создания собственных речевых движков; позволяет выбрать наиболее подходящий движок из доступных; позволяет обновлять или заменять модули, не переписывая код и не переустанавливая приложений. Отмечена перспектива использования в ТСР специальных языков разметки, таких как VXML и SSML.

Обоснована цель диссертационной работы и сформулированы основные задачи исследования.

Во второй главе диссертации рассмотрены аспекты использования речевых сигналов в ТСР незрячих и слабовидящих.

Исследованы строение речевого и слухового аппаратов человека, механизмы порождения и восприятия звуков речи, их характеристики.

Выделено 2 класса речевых сигналов (РС) – натуральные и ненатуральные. Натуральный РС – это речь, порождаемая непосредственно человеком. На звучание естественной речи влияют особенности строения голосового тракта и сенсомоторные навыки. Сложность точного и полного описания процесса порождения речи усугубляется сложностью её восприятия, обусловленной нелинейными свойствами слуха, поэтому весьма сложной задачей является модификация параметров РС без ущерба для его естественности. Ненатуральный РС – акустический сигнал, получаемый с помощью технических средств, и воспринимаемый человеком как речь. По способу получения выделены: предварительно записанный сигнал; сигнал, обработанный в реальном времени и синтезированный. Синтезированный сигнал, дифференцирован по технологии синтеза на артикуляторный, формантный и компилятивный РС. Среди систем синтеза речи наилучшей естественностью и разборчивостью обладают системы, использующие компилятивный метод, основанный на использовании образцов отдельных звуков.

Естественность речи определяется, главным образом, её просодией, или совокупностью ритмоинтонационных характеристик.

Разработана *информационная модель просодии*, включающая: *факторы, характеристики* и *параметры* просодии, а также *связи* между ними.

Установлено, что в большинстве систем синтеза речи по тексту при формировании просодических параметров в той или иной мере учитываются

факторы дикции и смысла, а эмоциональный фактор игнорируется ввиду сложности и малой изученности. Следовательно, для повышения качества синтезируемой речи необходимы исследования влияния эмоций на характеристики речи и разработка модели, описывающей проявление эмоций в РС.

В результате анализа различных психологических теорий, описывающих эмоции, установлено:

1. Эмоционально окрашенная речь (ЭОР) является результатом реализации экспрессивной функции эмоций.

2. В психологии эмоций большое распространение получила многомерная дискретная модель.

3. Классификация эмоций выполняется либо на основе набора признаков, либо по базовым эмоциям.

4. Параметром, присущим каждой эмоции, является сила её действия, определяемая на основе субъективных оценок.

Определены параметры для описания эмоций при построении модели их проявления в речевом сигнале, в том числе: *вид, сила, время возникновения* и *длительность проявления* эмоции.

Разработана методика формирования экспериментальной выборки образцов РС, создано множество «эмоциональных» образцов, включающее 50 фраз с выраженным эмоциональным окрасом, и множество «нейтральных» образцов, включающее 70 записей фраз без эмоционального окраса.

В качестве исследуемой эмоции выбрана эмоция радости.

Предложена методика пофонемной оценки образцов. Установлено, что эмоциональная характеристика слов и фраз является неравномерной и в большей мере определяется характеристикой фонем, находящихся ближе к концу слова, фразы или предложения.

Выдвинута гипотеза о локализации эмоциональной компоненты на тональных участках РС, т.е. на гласных и вокализованных согласных звуках. С целью проверки данной гипотезы проведён следующий эксперимент. Из полученного множества отобрано 20 образцов фраз и сформировано 10 пар, содержащих один эмоциональный (с баллом более 3,5) и один нейтральный образец. Образцы каждой пары содержали одинаковые фразы, произнесённые одним и тем же диктором, т.е. различались только эмоционально. Затем в нейтральных образцах была произведена замена сначала гласных, а затем и вокализованных согласных участков аналогичными, взятыми из эмоциональных образцов. Также проведён эксперимент по замене фрагментов эмоциональных образцов аналогичными, взятыми из нейтральных образцов. Полученные на каждом этапе образцы были сохранены и затем предъявлены экспертам. Оценки образцов, полученных заменой нейтральных фонем эмоциональными, представлены в таблице 1.

Замена эмоциональных фонем нейтральными имела обратный эффект: например, подстановка в эмоциональный образец с оценкой 5,0 гласных фонем из нейтрального образца с оценкой 1,4 понизила оценку первого образца до 2,4 , а последующая замена вокализованных – до 1,8.

Таблица 1.

Пара образов (диктор)	Средняя оценка уровня эмоций			
	«Нейтральный» образец	«Эмоциональный» образец	После замены гласных	После замены гласных и вокализованных
1. (ж)	1,2	4,8	3,8	4,2
2. (ж)	1,4	5,0	4,4	4,8
3. (м)	1,0	3,8	3,6	3,8
4. (м)	1,0	4,8	4,0	4,6
5. (м)	1,2	4,6	3,8	4,0

Полученный результат позволил в дальнейшем ограничиться рассмотрением гласных участков РС, поскольку параметры именно этих участков в основном определяют эмоциональный окрас речи.

Поскольку время действия эмоции значительно превосходит длительность фонемы, то для придания эмоционального окраса отдельной фонеме достаточно таких параметров, как *вид эмоции* и *сила эмоции*. При этом значение силы для данной фонемы должно быть рассчитано исходя из начального значения силы, интервала между моментом возникновения эмоции и началом звучания фонемы, а также заданной длительности проявления эмоции.

Для реализации синтеза ЭОР необходима модификация типовой схемы синтезатора таким образом, чтобы при формировании просодических характеристик РС учитывался и эмоциональный фактор.

Проведена оценка эмоциональной разборчивости зашумлённого РС. Сформировано множество образцов, содержащих по 5 фраз различной эмоциональности, при этом на каждый образец наложен белый шум определённой интенсивности. Образцы прослушаны 5 экспертами, ответившими затем на ряд вопросов.

Установлено, что благодаря особенностям слуховой системы человека, эмоциональная разборчивость речи значительно превышает разборчивость вербальную: эмоциональные различия между фразами фиксировались большинством экспертов при соотношении сигнал/шум, равном -10 дБ.

В третьей главе диссертации рассмотрены вопросы создания математического и алгоритмического обеспечения задачи автоматического анализа и коррекции эмоционального окраса РС.

Разработана сегментарная модель, позволяющая компактно и адекватно описывать вокализованные участки РС.

Локализация эмоциональной компоненты на гласных звуках, их небольшая длительность и квазипериодичность сделали перспективным использование разработку методов и алгоритмов анализа и модификации этих участков РС, основанных на работе с сигналом во временной области. Исходя из характерной формы РС на гласных и вокализованных участках, реализован следующий подход к построению модели:

1. Разбиение вокализованного участка на периоды основного тона (ОТ) с их последовательной нумерацией. Разработан и реализован алгоритм

автоматической разметки, включающий следующие шаги:

1.1. На осциллограмме образца выделяется вокализованный участок, границами которого выбираются точки локальных максимумов сигнала. Левая граница выделенного участка совпадает с левой границей p_0 начального периода ОТ, а правая граница участка – с правой границей p_n последнего периода ОТ.

1.2. Для временного ряда, представляющего собой значения оцифрованного речевого сигнала на выделенном участке, вычисляются оценки автокорреляционной функции (АКФ) при различных значениях лага:

$$R(\tau) = \frac{1}{N-\tau} \sum_{i=0}^{N-1-\tau} x(i) \cdot x(i+\tau), \qquad \tau \in [\tau_1; \tau_2]. \tag{1}$$

где: $R(\tau)$ – оценка АКФ, вычисленная для значения лага τ;

$x(i)$ – значение отсчёта i выделенного участка речевого сигнала;

$x(i+\tau)$ – значение отсчёта, сдвинутого относительно отсчёта i на лаг τ;

N – длина выделенного участка в отсчётах.

Значения τ_1 и τ_2 задаются исходя из отношения граничных значений периода ОТ к периоду квантования, или из обратного отношения соответствующих частот. Так, для мужского голоса, со значением частоты ОТ, лежащим в интервале 100-200 Гц, записанного с частотой квантования 22 050 Гц, граничными значениями будут: τ_1=22050/200≈110 ; τ_2=22050/100≈221.

1.3. При вычислении оценок АКФ фиксируется значение лага τ^*, при котором значение оценки $R(\tau)$ максимально:

$$R(\tau^*) = \max\{R(\tau_{\min}), R(\tau_{\min}+1), \ldots, R(\tau_{\max})\} \tag{2}$$

Правая граница периода ОТ устанавливается в точке локального максимума сигнала, принадлежащей окрестности δ, центр которой отстоит от левой границы периода на лаг τ^*. Таким образом, для правой границы периода ОТ справедливо условие:

$$x(p_{b+1}) = \max\{x(p_b + \tau^* - \frac{\delta}{2}), \ldots, x(p_b + \tau^* + \frac{\delta}{2})\}. \tag{3}$$

где: $x(p_{b+1})$ – значение сигнала в точке p_{b+1} ;

p_b – левая, а p_{b+1} – правая границы периода ОТ с номером b ;

τ^* – лаг, соответствующий максимальному значению оценки АКФ $R(\tau)$;

δ – размер окрестности для поиска локального максимума сигнала.

1.5. Операции по п.п. 1.2-1.4 выполняются для следующего периода.

2. Разметка каждого периода ОТ на *сегменты* – участки с одинаковым знаком приращения значения сигнала. Соответственно, границами сегментов являются точки изменения знака приращения: в дискретной последовательности отсчёт i со значением x_i является граничным, если

$$(x_{i-1} < x_i \geq x_{i+1}) \vee (x_{i-1} > x_i \leq x_{i+1}). \tag{4}$$

Сегменты каждого периода ОТ последовательно нумеруются от 0 до **S**-1.

3. Вычисление параметров сегментов. Если временной ряд, представляющий собой дискретную реализацию РС, соответствующего гласному зву-

ку, разбить граничными точками по условию (4), то в полученных в результате сегментах можно выделить общие признаки формы. Исходя из характерной формы большинства сегментов, для аппроксимации предложена функция

$$x_i = x_m + h \cdot \sin^k\left(\frac{\pi}{2} \cdot \frac{i-m}{l}\right), \quad (5)$$

где

$$h = x_n - x_m \quad (6)$$

– высота сегмента,

$$l = n - m \quad (7)$$

– его длительность.

В формулах (5)-(7) x_i – значение произвольного отсчёта ***i*** в сегменте, ограниченном отсчётами ***m*** и ***n*** со значениями x_m и x_n соответственно.

Таким образом, каждый сегмент характеризуется следующими параметрами: номером ***m*** и значением x_m начального отсчёта, длительностью ***l***, высотой ***h*** и коэффициентом формы ***k***.

Тогда задача аппроксимации сводится к нахождению по известным значениям отсчётов РС параметров сегментов, позволяющих с заданной точностью представить сигнал.

Аппроксимирующая функция (5) позволяет кодировать вокализованные участки РС набором параметров сегментов, которые вычисляются исходя из имеющихся значений отсчётов сигнала, а также декодировать сигнал, вычислив значения отсчётов по заданным параметрам сегментов.

Информационная модель просодии РС, предложенная во 2-й главе, расширена параметрами сегментарной модели. На рисунке 1 показана взаимосвязь сегментарной модели РС и информационной модели его просодии.

Разработан алгоритм расчёта параметров сегментарной модели.

Предложен способ шаблонного представления, позволивший решить проблему непостоянного количества сегментов и описывать динамику параметров конкретного сегмента в последовательности периодов основного тона, составляющих вокализованный участок РС.

Шаблон – аппроксимация сигнала, при которой количество сегментов в каждом периоде ОТ постоянно и равно заданному *размеру шаблона* ***S'***.

Построение шаблона состоит в уменьшении количества границ сегментов в каждом периоде ОТ до значения, равного заданному размеру шаблона ***S'***, путём поиска сегментов с минимальной высотой и объединения их с соседними. Целью этой операции – получение *карты принадлежности* – таблицы, отражающей, в состав какого шаблонного сегмента входит данный нативный сегмент данного периода ОТ.

Установлена зависимость между средними значениями параметров шаблонных сегментов и эмоциональным состоянием диктора, позволяющая использовать средние значения в качестве признака эмоциональности РС.

Выполнен кластерный анализ параметров шаблонных сегментов, вычисленных для фонем различной эмоциональности. Для анализа были ото-

браны фонемы с различными оценками эмоциональности: для нейтральных фонем средние (по экспертам) значения оценок лежали в интервале [1,0; 2,0], а для эмоциональных - в интервале [4,0; 5,0], дисперсия оценок не более 0,2. В фонемах рассматривались 5 начальных периодов ОТ, для которых были вычислены 8-сегментные шаблоны, то есть каждая фонема описывалась параметрами 40 шаблонных сегментов (ШС). Установлено, что наибольшая нестационарность характерна для длительности ШС ***l***. В данном случае анализ проводился по длинам первых трёх ШС каждого периода ОТ, таким образом, каждая фонема как объект анализа описывалась 15 признаками. После выделения из этих 15-и 4-х наиболее значимых признаков сформирована обучающая выборка ОВ1, состоящая из 20 нейтральных и 37 эмоциональных объектов. Выполнена кластеризация выборки $ОВ_1$ методом Варда. При разделении на 5 кластеров допущена 1 ошибка. При разделении на 2 кластера допущено 3 ошибки. Удаление неверно классифицированных объектов из выборки и её повторное разделение на 2 кластера выполнено без ошибок.

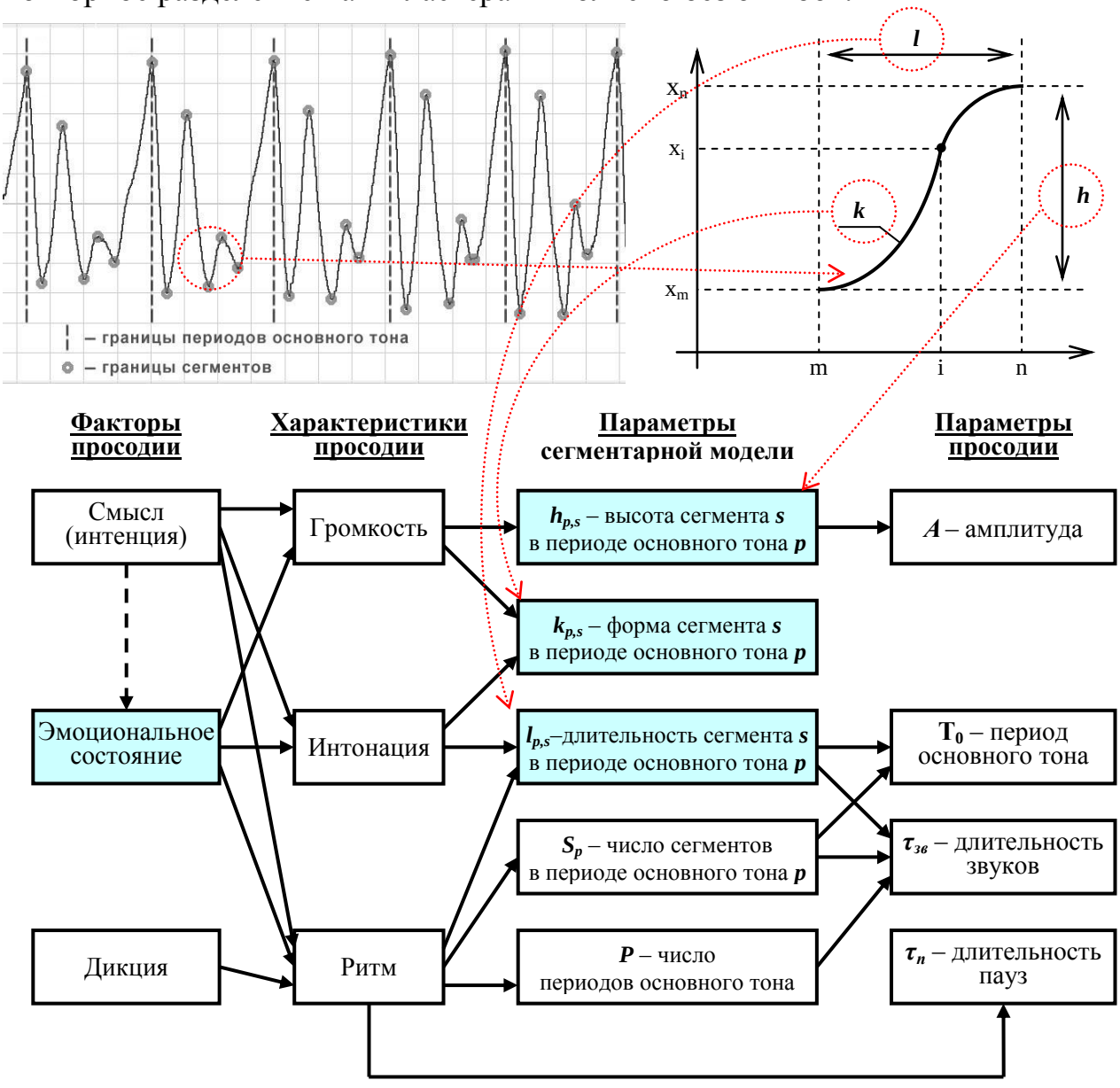

Рис. 1. Интеграция параметров сегментов в информационную модель просодии РС.

Кластеризация показала, что имеет место корреляция между объективной (измеряемой длиной сегментов шаблона) и субъективной (воспринимаемой на слух эмоциональностью) характеристиками речевого сигнала. Это указывает на принципиальную возможность автоматического анализа и коррекции эмоционального окраса РС в модуле синтеза речи по тексту системы реабилитации слабовидящих.

Разработан и опробован алгоритм *MSH-коррекции*, т.е. коррекции просодических характеристик РС (прежде всего, эмоционального окраса и интонации) на основе изменения отношений параметров в текущем периоде ОТ к значениям в предыдущем периоде ОТ:

$$dl_{p,s} = \frac{l_{p,s}}{l_{p-1,s}}, \tag{12}$$

$$dh_{p,s} = \frac{h_{p,s}}{h_{p-1,s}}, \tag{13}$$

В формулах (12)-(13) индекс p=[1; P-1] указывает на номер периода ОТ, а индекс s=[0; S'-1] на номер сегмента в шаблоне (здесь P – количество сегментируемых периодов ОТ, а S' – размер шаблона).

Алгоритм MSH-коррекции. Пусть имеется два образца фонемы, обладающих различной степенью эмоционального окраса. Если образец $\boldsymbol{A}$ имеет оценку эмоциональности $\boldsymbol{e_A}$, а оценка образца $\boldsymbol{B}$ равна $\boldsymbol{e_B}$, и требуется скорректировать значение $\boldsymbol{e_B}$ так, чтобы сделать его равным или максимально близким значению $\boldsymbol{e_A}$, то выполняются следующие операции:

1. Открытие wav-файла «эталонного» образца $\boldsymbol{A}$.
2. Сегментация образца $\boldsymbol{A}$.
3. Вычисление шаблона размером $\boldsymbol{S'}$ сегментированного образца $\boldsymbol{A}$.
4. Вычисление матриц отношений ШС образца $\boldsymbol{A}$.
5. Сохранение матриц отношений ШС образца $\boldsymbol{A}$ в msh-файл.
5. Открытие wav-файла корректируемого образца $\boldsymbol{B}$.
6. Сегментация образца $\boldsymbol{B}$.
7. Вычисление шаблона размером $\boldsymbol{S'}$ сегментированного образца $\boldsymbol{B}$.
8. Вычисление матриц отношений ШС образца $\boldsymbol{B}$.
9. Загрузка матриц отношений ШС образца $\boldsymbol{A}$ из msh-файла.
10. Коррекция параметров нативных сегментов образца $\boldsymbol{B}$ исходя из соотношения загруженных («эталонных») и вычисленных значений отношений шаблонных сегментов. Если при построении шаблона в периоде основного тона $\boldsymbol{p}$ нативный сегмент $\boldsymbol{x}$ вошёл в состав шаблонного сегмента $\boldsymbol{z}$, то скорректированные значения параметров сегмента вычисляются по формулам:

$$l^C_{p,z} = l^C_{p-1,z} \cdot \sqrt{dl^A_{p,z} \cdot dl^B_{p,z}} \quad , l^C_{p=0,z} = l^B_{p=0,z}\,, \tag{14}$$

$$h^{C}_{p,z} = h^{C}_{p-1,z} \cdot \sqrt{dh^{A}_{p,z} \cdot dh^{B}_{p,z}}\ , h^{C}_{p=0,z} = h^{B}_{p=0,z}\ , \qquad (15)$$

где: $l^{A}_{p,x}$ – длительность нативного сегмента *x* в периоде основного тона ***p*** образца ***A***; $h^{B}_{p,z}$ – высота шаблонного сегмента ***z*** в периоде основного тона ***p*** образца ***B***.

11. Коррекция значений отсчётов wav-файла образца ***B*** в соответствии с формулой (5) и новыми значениями нативных сегментов.

12. Сохранение wav-файла образца ***B***.

Таким образом, коррекция длительности и высоты нативных сегментов заключается в умножении текущего значения параметра на коэффициент, равный отношению нового (эталонного, загружаемого) и старого (вычисляемого) значений параметра шаблонного сегмента, к которому принадлежит данный нативный сегмент. Значения коэффициентов формы *k* не корректируются. Это позволяет сохранять индивидуальные особенности, такие, как тембр, характерный для данного диктора, и одновременно изменять просодические характеристики, в т.ч. эмоциональный окрас воспроизводимого звука.

Экспериментально определены наилучшие для выполнения MSH-коррекции размеры шаблонов большинства гласных фонем: для фонем [и], [о], [у] это 8 сегментов, для фонем [а], [е] – 10.

Выполнена коррекция нескольких фраз, синтезированных TTS-модулем "Digalo" (диктор Nicolai), позволившая улучшить естественность их звучания с 3,0 до 4,2 баллов по пятибалльной шкале (по оценке экспертов).

Проведён кластерный анализ выборки $ОВ_2$, сформированной аналогично $ОВ_1$, из 16 исходных (синтезированных нейтральных), 28 скорректированных (синтезированных) и 41 эталонных (записанных эмоциональных) объектов. Разбиение $ОВ_2$ на 2 кластера по методу Варда показало, что скорректированные по MSH-алгоритму фонемы по длинам шаблонных сегментов объединяются в один кластер с эмоциональными, а нейтральные образуют отдельный от них кластер. Это позволяет сделать вывод о том, что синтезированные фонемы, параметры которых скорректированы MSH-алгоритмом, по своим характеристикам близки к фонемам естественного эмоционального РС.

В четвертой главе диссертации рассмотрены вопросы реализации разработанных моделей и алгоритмов в ТСР незрячих и слабовидящих.

Предложена архитектура программной системы реабилитации (рисунок 2), включающая следующие компоненты:

1. Активные приложения – программы, запускаемые пользователем, и предназначенные для решения различных задач (просмотр веб-страниц, работа с электронной почтой, редактирование документов, и т.д.)

2. Программа «Голосовой менеджер», предназначенная для отслеживания действий пользователя и инициируемых ими событий в активных приложениях, подготовка сообщений, содержащих информацию о действиях и событиях, а также сообщений, повторяющих содержание открытых документов, и передача этих сообщений речевому движку для их озвучивания.

3. Речевой движок, преобразующий получаемые в виде простого или содержащего SSML-разметку текста сообщения в речевой сигнал.

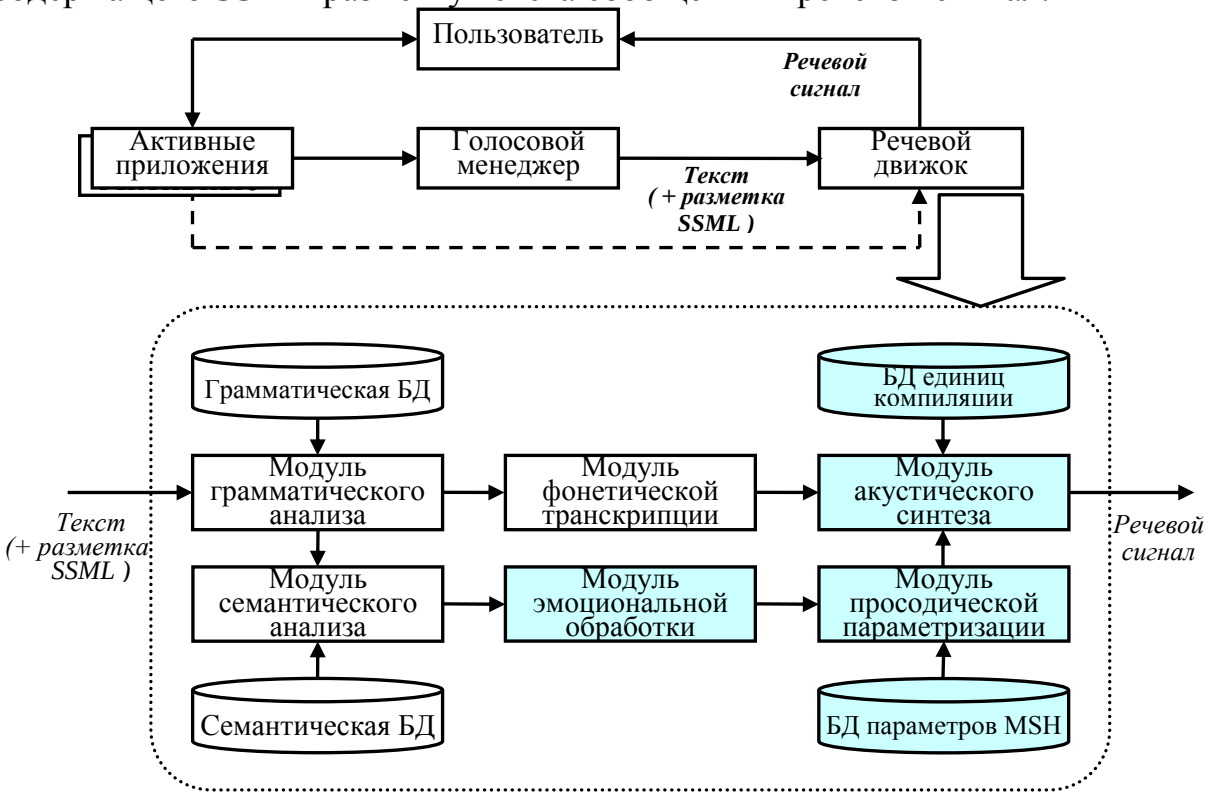

Рис. 2. Система реабилитации на основе модуля «Голосовой менеджер» и речевого движка, реализующего синтез эмоционально окрашенной речи.

Структура «Голосового менеджера» включает следующие модули:

1. *Модуль обработки событий приложений* – содержит функции-обработчики событий, происходящих в активных приложениях; нужен для отслеживания процессов в системе и действий пользователя. В процессе обработки любого события приложения проверяется: а) требует ли событие генерации речевого сообщения пользователю; б) требует ли события выполнения пользователем какого-либо действия. Если событие требует сообщения, то происходит его генерация в соответствующем блоке и передача речевому движку. При этом необходимость генерации сообщения оценивается исходя из данных о квалификации пользователя и степени его патологии, содержащихся в настройках его профиля.

2. *Модуль генерации событий для приложений* – является «исполнительным механизмом» системы реабилитации. В ответ на определённые действия пользователя в эмулирует «работу нормального пользователя».

3. *Модуль генерации сообщений для пользователя* – выполняет подготовку текстовых сообщений, подлежащих чтению вслух речевым движком. Материалом для сообщений служит содержимое текстовых документов, а также текстовое содержимое окон и элементов управления.

4. *База данных профилей настроек* содержит сведения о том, какую информацию и в каком объёме следует озвучивать при работе конкретного пользователя. Настройки задаются исходя из степени патологии, уровня квалификации и психологических особенностей пользователя.

5. *Модуль управления* – координирует функционирование всех модулей в процессе работы приложения «Голосовой менеджер».

Разработана структура речевого движка, позволяющая синтезировать естественно звучащую речь, используя модель эмоционального настроя диктора и алгоритм MSH-коррекции просодических характеристик РС.

Основным отличием предложенной структуры речевого движка от традиционной является наличие модуля эмоциональной обработки (МЭО).

Задача МЭО – эмоциональная разметка высказывания, которая может формироваться следующим образом. В модуле можно реализовать некий «эмулятор настроения», задача которого – моделирование и постоянная индикация эмоционального состояния виртуального диктора. Последнее может фиксироваться в наборе переменных, каждая из которых хранит степень проявления определённой эмоции. При первом запуске TTS-движка происходит инициализация переменных некоторыми значениями по умолчанию.

Настроение диктора, используемого движком, отражает индекс текущего настроения. Изменение индекса происходит на величину, рассчитываемой на основе множества факторов. Эта величина используется для аддитивной коррекции текущего настроения, или эмоционального статуса диктора.

В соответствии с текущим эмоциональным состоянием диктора выполняется пофонемная эмоциональная SSML-разметка текста. Она заключается в маркировке гласных фонем во фразах озвучиваемого текста специальными дополнительными SSML-тэгами. Предлагается ввести тэг <*emotion*> с атрибутами *type* – вид эмоции и *volume* – сила эмоции.

Разработаны форматы файлов для хранения результатов сегментации, параметров шаблонов и их приращений. Метод анализа и коррекции эмоционального окраса РС на основе разработанной сегментарной модели гласных и вокализованных звуков реализован программно в виде набора классов функций, позволяющих работать со следующими представлениями сигнала: **WAV** – стандартный формат звуковых файлов в ОС Windows (в данной работе использовался режим моно, 22050 Гц,16 бит); **SEG** – формат для сохранения результатов сегментации WAV-данных; **SHA** – формат для сохранения параметров сегментов шаблонов; **SSH** – формат для сохранения средних значений параметров шаблонных сегментов; **MSH** – формат для сохранения отношений параметров шаблонных сегментов, рассчитанных по формулам (12) и (13).

Создано программное обеспечение, позволяющее анализировать и модифицировать РС путём вычисления и редактирования параметров сегментов. Разработка защищена свидетельством Роспатента на программу для ЭВМ.

SSML-разметка текста используется при формировании просодических характеристик высказывания. Просодию гласных фонем вместо традиционных параметров (амплитуда A, длительность фонемы τ, ЧОТ F_0) предлагается описывать шаблонными приращениями (dL, dH, dK).

Соответственно, БД единиц компиляции должна содержать не WAV-представление гласных звуков, а параметры сегментов (матрицы длительностей $L=[l_{ps}]$, высот $H=[h_{ps}]$ и коэффициентов формы $K=[k_{ps}]$). Также необхо-

дима БД MSH-параметров гласных фонем, соответствующих различным по типу и силе эмоциям.

ЗАКЛЮЧЕНИЕ

1. Выполнен анализ ТСР незрячих и слабовидящих, определена роль речевого синтеза. Установлено, что применение синтеза речи в ТСР сдерживается недостаточным качеством получаемого сигнала. Синтезированный сигнал, имея хорошую разборчивость, звучит недостаточно естественно, что обусловлено его неадекватной просодией.

2. Разработана информационная модель просодии РС, включающая факторы, характеристики, параметры и связи между ними. Основными факторами просодии являются смысл, вкладываемый в высказывание говорящим, его дикция и эмоциональное состояние. Установлено, что в большинстве систем синтеза речи по тексту при формировании просодических параметров в той или иной мере учитываются факторы дикции и смысла, а эмоциональный фактор игнорируется ввиду сложности и малой изученности. Поэтому для повышения качества синтезируемой речи путём улучшения её естественности требуется исследование влияния эмоций на характеристики речи и разработка модели, описывающей эмоциональные проявления в РС.

3. Проведены исследования эмоциональных проявлений в РС. Разработана и реализована методики получения и экспертной оценки образцов РС, обладающих различной эмоциональной характеристикой. Выдвинута и экспериментально подтверждена гипотеза о локализации эмоциональной компоненты на гласных и отчасти на вокализованных звуках РС. Разработана методика пофонемной оценки образцов РС.

4. Разработана сегментарная модель представления вокализованных участков РС, позволяющая компактно и адекватно описывать гласные и вокализованные согласные во временной области. Параметры сегментарной модели интегрированы в модель просодии, что позволяет, изменяя параметры сегментов, управлять просодией РС, в том числе его эмоциональным окрасом.

5. Предложен способ шаблонного представления, позволяющий описывать динамику параметров сегментов в последовательности периодов основного тона, составляющих гласный или вокализованный участок РС. Установлена зависимость между средними значениями параметров шаблонных сегментов и эмоциональным состоянием диктора, позволяющая использовать средние значения в качестве признака эмоциональности РС.

6. Разработан и опробован алгоритм коррекции просодических характеристик РС (прежде всего, эмоционального окраса и интонации) на основе изменения отношений параметров шаблонных сегментов (MSH-коррекция). Найдены наилучшие для выполнения MSH-коррекции размеры шаблонов большинства гласных фонем. Достоверность алгоритма MSH-коррекции подтверждена результатами кластерного анализа выборки, включающей исходные (нейтральные), эталонные (эмоциональные) скорректированные (с возросшей в результате эмоциональностью) образцы речевого сигнала.

7. Создано ПО для анализа и модификации РС путём вычисления и ре-

дактирования параметров сегментов. Разработаны форматы файлов для хранения результатов сегментации, параметров шаблонов, их приращений и отношений. Созданы свободно распространяемые классы функций для преобразования РС из WAV-формата в форматы SEG, SHA, ASH, MSH и обратно.

8. Разработана архитектура программной системы реабилитации слабовидящих, включающая модуль «Голосовой менеджер» и речевой движок, позволяющий синтезировать естественно звучащую речь, используя модель эмоционального настроя диктора и алгоритм MSH-коррекции просодических характеристик РС. Система позволяет незрячим и слабовидящим пользователям работать на персональном компьютере с любыми приложениями, имеющими стандартный программный интерфейс, не требуя специальных дорогостоящих устройств. Это делает работу с компьютером доступной для широкого круга пользователей с различными патологиями зрения.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1. Калюжный, М.В. Коррекция просодических характеристик речевого сигнала в средствах реабилитации незрячих и слабовидящих [Текст] / М.В. Калюжный, Н.Н. Филатова // Информационно-управляющие системы. – 2008. №1. – СПб.: РИЦ ГУАП, 2008. – С. 54–57.

Другие статьи и материалы конференций:

2. Калюжный, М.В. Программа для расчёта параметров сегментарной модели речевого сигнала [Текст] / М.В. Калюжный, Н.Н. Филатова // НАУЧНАЯ СЕССИЯ МИФИ-2008. Сборник научных трудов. В 15 томах. Т. 12. Информатика и процессы управления. Компьютерные системы и технологии. – М.:МИФИ, 2008. – С. 50–51.

3. Калюжный, М.В. Анализ параметров сегментарной модели речевого сигнала [Текст] / М.В. Калюжный // Математические методы в технике и технологиях – ММТТ-21. Сборник трудов. XXI Международной научной конференции в 10 томах. Т. 9. – Саратов: СГТУ, 2008. – С. 65–66.

4. Калюжный М.В. Моделирование эмоциональных проявлений в речевом сигнале //Свидетельство об официальной регистрации программы для ЭВМ № 2007614294. – М.: Роспатент, 2007.

5. Калюжный, М.В. Параметрическое описание речевого сигнала в модели эмоционально окрашенной речи [Текст] / М.В. Калюжный, Н.Н. Филатова // Электроника и информатика – 2005. V Международная научно-техническая конференция: Материалы конференции: в 2 ч. – М.: МИЭТ, 2005. – С. 11–12.

6. Калюжный, М.В. Исследование проявлений эмоций в речевом сигнале [Текст] / М.В. Калюжный // Вестник Тверского государственного технического университета: Научный журнал – Тверь, 2005. – Вып. 7. – С. 102–106.

7. Калюжный, М.В. Синтез естественно звучащей речи на основе модели ЭОР [Текст] / М.В. Калюжный, Н.Н. Филатова // Компьютерные технологии в управлении и диагностике: сб. научн. тр. / Тверской гос. тех. ун-т. – Тверь, 2004. – С. 101-104.