УДК 007:681

КОРРЕКЦИЯ ПРОСОДИЧЕСКИХ ХАРАКТЕРИСТИК РЕЧЕВОГО СИГНАЛА В СРЕДСТВАХ РЕАБИЛИТАЦИИ НЕЗРЯЧИХ И СЛАБОВИДЯЩИХ

М. В. Калюжный,
ассистент
Н. Н. Филатова,
доктор техн. наук, профессор
Тверской государственный технический университет

Рассмотрены аспекты применения средств речевого синтеза в системах реабилитации незрячих и слабовидящих. Описаны исследования проявлений эмоций в голосе. Предложена модель речевого сигнала, позволяющая анализировать и изменять эмоциональный окрас речи с целью улучшения ее естественности.

Анализ современных систем реабилитации незрячих и слабовидящих позволяет выделить два основных подхода: *коррекционный*, состоящий в восстановлении, коррекции или поддержании на приемлемом уровне функций, утраченных либо ослабленных ввиду патологии, и *компенсационный*, подразумевающий использование возможностей других функциональных систем организма для компенсации функциональной недостаточности пораженной системы.

Первый подход реализуется с помощью технических средств реабилитации (ТСР), позволяющих увеличить резкость, размер или контрастность изображения. К таким средствам относятся очки, контактные линзы, оптические увеличители. Второй подход подразумевает использование других каналов восприятия (осязания и слуха) и построен на применении рельефных изображений, а также акустических сигналов, главным образом, речи. Распространение ТСР, использующих тактильный ввод/вывод информации, ограничивают достаточно высокие требования к навыкам работы с ними и их высокая стоимость (2000–4500 $) [1].

Широкое применение синтеза речи в средствах реабилитации незрячих и слабовидящих сдерживается недостаточным качеством получаемого речевого сигнала (РС). Качество речи определяют такие ее характеристики как разборчивость и естественность. Среди систем синтеза речи наилучшие характеристики обеспечивают системы, использующие компилятивный метод синтеза.

Звучание речи зависит от просодических характеристик, определяемых:

1) индивидуальными особенностями артикуляции и фонации (дикцией);

2) смыслом, который вкладывается в высказывание говорящим (диктором);

3) эмоциональным состоянием диктора.

При формировании синтезированной речи фактор смысла учитывается, исходя из возможностей модуля лингвистического анализа текста. Индивидуальные особенности артикуляции учитываются всегда, поскольку синтезатор воспроизводит дикцию человека, голос которого использовался при формировании базы элементов компиляции. Однако эмоциональный фактор, как правило, не учитывается в связи с чрезвычайной сложностью задачи. Речь, синтезированная на основе неполных просодических характеристик, обладает нейтральными интонациями без всяких эмоций, что затрудняет ее длительное восприятие.

Таким образом, одной из главных задач построения речевого интерфейса с хорошими эргономическими свойствами для систем реабилитации незрячих и слабовидящих является исследование и моделирование эмоционально окрашенной речи (ЭОР).

Информационная модель просодии РС, включающая факторы, характеристики, параметры и связи между ними, предложена в работе [2]. Для ЭОР характерна специфическая просодия, т. е. определенное сочетание громкости, тембра, интонации и ритма. Важнейшей просодической характеристикой речевого сигнала является частота основного тона (ЧОТ), которая с вероятностью 0,95 составляет 100–200 Гц у мужских голосов и 220–350 Гц у женских [3]. Основной тон у одного и того же человека может значительно меняться в зави-

симости от ситуации и эмоционального состояния. Динамика изменения основного тона определяет интонации: в русской речи для интонационно нейтральных предложений характерно плавное понижение ЧОТ к концу предложения, а для вопросительных — повышение. Отмечена тесная взаимосвязь эмоций и интонаций в речевом сигнале [4]. Поскольку интонации определяются изменением ЧОТ в процессе произнесения слов и фраз [3], а само понятие ЧОТ применимо только к гласным и вокализованным согласным, логично полагать, что имеет место локализация эмоциональной компоненты на гласных и вокализованных участках РС.

В целях проверки данной гипотезы проведен следующий эксперимент. Нескольким дикторам (трем мужчинам и двум женщинам) было предложено прочесть перед микрофоном текст юмористического характера, способный с большой вероятностью вызвать эмоцию радости. Спустя некоторое время дикторы повторяли фрагменты текста, вызвавшие эмоции при первом прочтении. Записи голосов дикторов сохранялись в аудиофайлах, а затем из них вырезались фрагменты и сохранялись в виде отдельных файлов образцов. Образцы с одинаковым содержанием объединялись в группы. Сформировано множество L из 30 образцов, произнесенных с различной степенью эмоционального окраса. Полученные образцы были предъявлены пяти экспертам, оценивавшим степень проявления эмоции в каждом образце по пятибалльной шкале, затем оценки каждого образца усреднялись. Таким образом, были выделены 9 эмоциональных (получивших более 3,5 балла), 13 умеренных (от 1,5 до 3,5 балла) и 8 нейтральных (от 0 до 1,5 балла) образцов. Образцы из L разбиты на пары, каждая из которых включает записи одинаковой фразы, произнесенной одним и тем же диктором с эмоциональным окрасом и без (нейтральный образец). В нейтральных образцах была произведена замена сначала гласных, а затем и вокализованных согласных аналогичными участками, взятыми из эмоциональных образцов. Был проведен и обратный эксперимент, модифицированные образцы были предъявлены экспертам (табл. 1, 2).

■ ***Таблица 1***

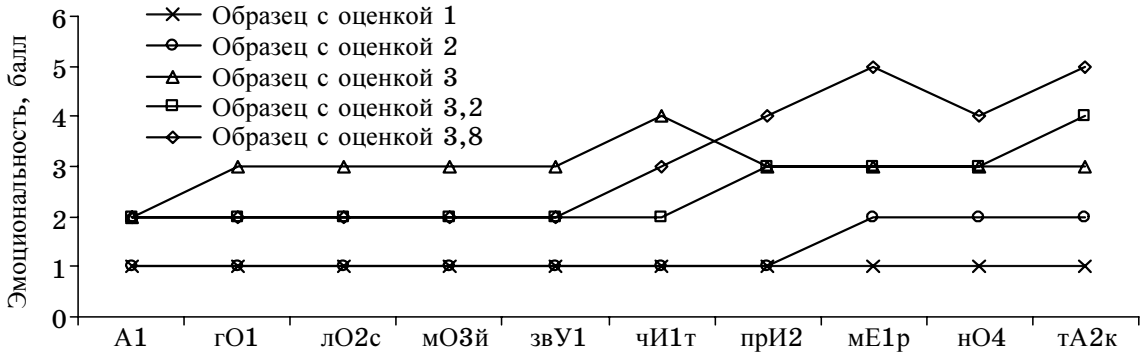
№ пары образцов. Диктор	Средняя оценка уровня эмоций при замене нейтральных фрагментов эмоциональными, балл			
	«Нейтральный» образец	«Эмоциональный» образец	Замена гласных	Замена гласных и вокализованных
1. Ж	0,2	4,8	3,8	4,2
2. Ж	1,4	5,0	4,4	4,8
3. М	0,0	3,8	3,6	3,8
4. М	1,0	4,8	4,0	4,6
5. М	1,2	4,6	3,8	4,0

■ ***Таблица 2***

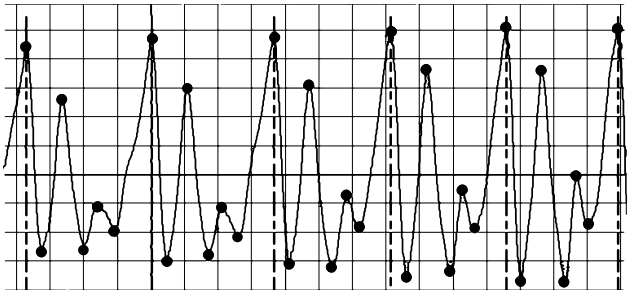
№ пары образцов. Диктор	Средняя оценка уровня эмоций при замене эмоциональных фрагментов нейтральными, балл			
	«Эмоциональный» образец	«Нейтральный» образец	Замена гласных	Замена гласных и вокализованных
1. Ж	4,8	0,2	1,0	0,6
2. Ж	5,0	1,4	2,4	1,8
3. М	3,8	0,0	0,8	0,2
4. М	4,8	1,0	2,2	1,4
5. М	4,6	1,2	2,0	1,4

Эксперты отметили, что в образцах, получивших равные или близкие оценки, распределение эмоций на протяжении фразы или даже слова неравномерно и уникально. Это потребовало проведения *пофонемного* анализа, когда эксперт, прослушивая образец, строил временную диаграмму эмоционального состояния диктора в процессе произнесения фразы (рис. 1).

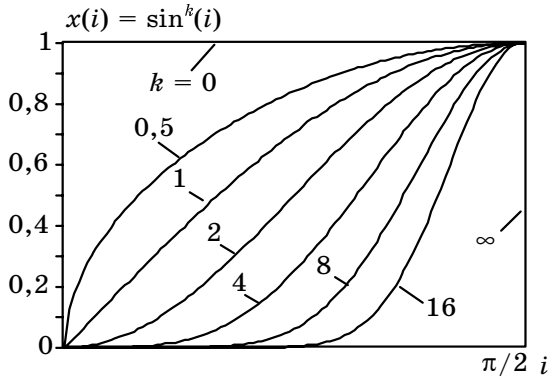
Разработана модель эмоционально окрашенных гласных и вокализованных согласных [5]. Модель построена на представлении указанных участков как сигналов с повторяющимися признаками формы. Предложенный подход заключается в разметке гласных и вокализованных участков РС сначала на периоды ОТ, а затем на сегменты —



■ ***Рис. 1. Оценка образцов*** (*фраза «А голос мой звучит примерно так»*)



■ ***Рис. 2.*** *Разметка фонемы «О» на периоды основного тона и сегменты: ¦ — границы периодов основного тона; • — границы сегментов*



■ ***Рис. 4.*** *Форма сегментов с различными значениями k*

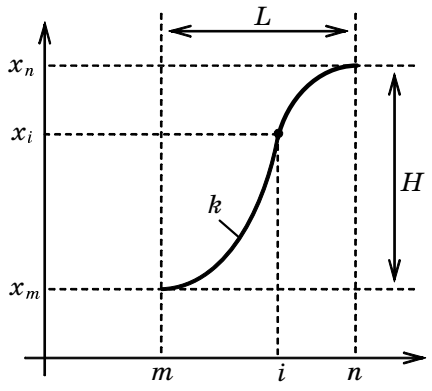
участки с одинаковым знаком приращения сигнала (рис. 2).

Исходя из характерной формы сегментов для прогноза значения x_i произвольного отсчета i сигнала внутри сегмента (рис. 3), ограниченного отсчетами m и n со значениями x_m и x_n соответственно, предложена функция

$$x_i = x_m + H \sin^k\left(\frac{\pi}{2} \cdot \frac{i-m}{L}\right).$$

В этом случае каждый сегмент характеризуется следующими параметрами: номером m и значением x_m начального отсчета, длительностью $L = n - m$, высотой $H = x_m - x_n$ и коэффициентом формы k. Варьирование k позволяет получить спрогнозированную форму сегмента, наиболее близкую к исходному сигналу (рис. 4).

Предлагаемая концепция, вполне согласуясь с доминирующим в настоящее время подходом к построению систем речевого синтеза, требует включения в типовую структуру синтезатора дополнительного модуля настройки эмоционального окраса РС. Решение задачи «эмоциональной коррекции» исходного текста возможно путем расширения функций модуля лингвистического анализа или предварительной разметки текста с помощью специальных языков (VXML и SSML).



■ ***Рис. 3.*** *Аппроксимация сигнала сегментом с параметрами L, H, k*

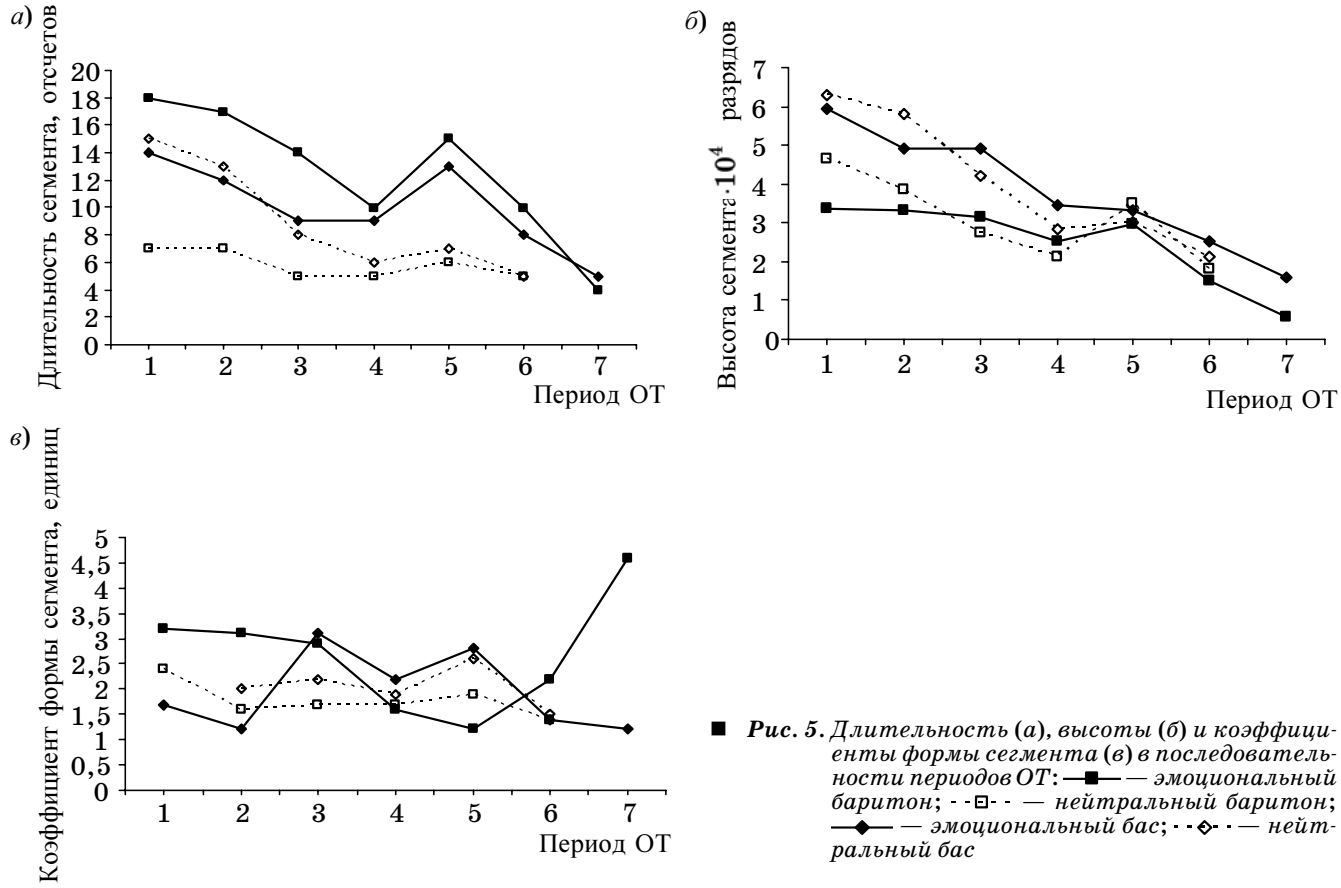
Для анализа динамики параметров сегментов в последовательностях периодов ОТ по описанной выше методике, но с участием большего числа дикторов, создано множество образцов РС. В роли дикторов выступали добровольцы в возрасте 20–30 лет, обладающие правильной дикцией и средним (80–120 слов в минуту) темпом речи, обеспечивающими разборчивость, близкую к 100 %, а также определенной (воспринимаемой на слух) эмоциональной выразительностью голоса, позволяющей проводить экспертную оценку образцов. Запись проводилась с помощью микрофона, подключенного к звуковой карте Creative SB AWE64. Все образцы записаны в формате Windows PCM (wav) с частотой дискретизации 22 050 Гц и разрядностью 16 бит. У образцов, принадлежащих одному диктору, но различных по интонационной и эмоциональной характеристикам, наблюдается сходство формы и размеров сегментов, находящихся в соседних периодах ОТ. Вместе с тем *изменения* формы и размеров сегментов носят общий характер у образцов, принадлежащих разным дикторам, но имеющих схожие интонацию и эмоциональный окрас. Графики на рис. 5, *а—в* иллюстрируют изменение параметров соответствующих сегментов в последовательности периодов ОТ, составляющих образцы.

Проведенные исследования позволяют сделать следующие выводы.

1. Экспериментально доказано, что локализация эмоциональной компоненты наблюдается на гласных и отчасти на вокализованных согласных звуках. При построении модели эмоционально окрашенной речи целесообразно ограничиться рассмотрением гласных участков РС.

2. Анализ временных диаграмм позволяет утверждать, что на эмоциональную оценку фразы в большей степени влияет оценка фонем, расположенных ближе к концу фразы.

3. Проявление эмоций (на примере радости) имеет общие закономерности в голосах различного тембра: изменения формы и размеров сегментов носят общий характер у образцов, принадлежащих разным дикторам, но имеющих схожие интонацию и эмоциональный окрас.



■ ***Рис. 5.*** *Длительность (а), высоты (б) и коэффициенты формы сегмента (в) в последовательности периодов ОТ: —■— — эмоциональный баритон; - -□- - — нейтральный баритон; —◆— — эмоциональный бас; - -◇- - — нейтральный бас*

4. Результатом работы явилась модель представления гласных и вокализованных участков РС, изменение параметров которой позволяет модифицировать просодические характеристики речи и имитировать проявление различных эмоций в голосе. Разработка алгоритмов модификации параметров модели сделает возможным синтез эмоционально окрашенной речи, близкой по звучанию к естественной, что позволит повысить качество систем реабилитации незрячих и слабовидящих.

Литература

1. http://www.baum.de/de/produkte/vario40.php
2. **Калюжный М. В.** Исследование проявлений эмоций в речевом сигнале // Вестник Тверского государственного технического университета. Тверь, 2005. С. 102–106.
3. **Секунов Н. Ю.** Обработка звука на РС. СПб.: БХВ-Петербург, 2001. 1248 с.
4. **Долотин К. И.** Эмоциональная интонация: проблема контекстуальной обусловленности признаков. 2000. http://www.philol.msu.ru/rus/gorn/arso/dolot.htm
5. **Калюжный М. В., Филатова Н. Н.** Параметрическое описание речевого сигнала в модели эмоционально окрашенной речи // Электроника и информатика — 2005. V Междунар. науч.-техн. конф.: Материалы конф. Ч. 2. М.: МИЭТ, 2005. С. 11–12.