УДК 004.934

АНАЛИЗ ПРИЗНАКОВ ЭМОЦИОНАЛЬНО ОКРАШЕННОЙ РЕЧИ

К.В. Сидоров, Н.Н. Филатова

В настоящее время пристальное внимание уделяется вопросам обработки информации и принятия решений при человеко-компьютерном взаимодействии. Эффективность данного процесса во многом зависит от качества информации, поступающей от пользователя автоматизированной системы и целенаправленности воздействия человека на объекты исследования. Достижение цели диалогового взаимодействия ЭВМ и пользователя возможно при учете большинства аспектов, характеризующих речевые потоки, возникающие в процессе общения.

Труд человека в системах управления техникой (деятельность человека-оператора) связан с периодическим, иногда довольно длительным и интенсивным воздействием экстремальных значений профессиональных, социальных, экологических факторов, которое сопровождается эмоциями, перенапряжением физических и психических функций, деструкцией деятельности.

Одним из источников эмоций является речевой сигнал (РС). Русский язык содержит около 40% эмоционально окрашенных слов. Эмоции кодируются определенными акустическими параметрами в РС, понимание этих особенностей акустического кодирования эмоций позволит понять сам механизм восприятия эмоций и их выражения [3].

Исследования РС проводились многими учеными с целью описания как лингвистических, так и технических характеристик речи. Большой вклад в развитие науки в области речевой акустики внесли ученые: Г. Фант, Дж. Фланаган, М.А. Сапожков, В.Н. Сорокин, В.И. Галунов, Б.М. Лобанов, Т.К. Винцюк, Н.В. Витт, Л.В. Златоустова, А.В. Аграновский, Н.Г. Загоруйко, Р.К. Потапова, Ю.А. Косарев, А.Л. Ронжин, М.В. Хитров, В.К. Иоффе, С.Л. Коваль, В.Г. Михайлов, В.П. Бондаренко, Л.Н. Балацкая, Е.Л. Чойнзонов и другие [1, 5].

Эмоционально окрашенная речь (ЭОР) находит применение во многих сферах жизнедеятельности человека и является востребованной функцией в современных автоматизированных системах управления, реабилитации и протезирования, срочного оповещения и т.п. В последние годы явно усилился интерес к анализу РС как объективного показателя эмоционального состояния (ЭС) человека, выполняющего ответственную деятельность космонавта, летчика, оператора АЭС, диспетчера аэропорта и т.д. (Лукьянов, Фролов, 1969; Таубкин, 1977; Williams, Stevens, 1969, 1972; Older, Jenney, 1975; Kuroda, Fujiwara, Okamura, Utsuki, 1976; Congleton, Jones, Shiflett и др., 1997; Rothkrantz, Wiggers и др., 2004; Sigmund, 2004; Хроматиди, 2005; Airas, Alku, 2006; Johannes, Wittels и др., 2007; Соловьева, 2008; Chen, 2008; Siging, 2009; Розалиев, 2009; Калюжный, 2009; Перервенко, 2009; Morist, 2010). Однако, несмотря на множество исследований в данной области, проблема автоматического распознавания ЭС говорящего по речи на данный момент не является полностью решенной, в частности, отсутствует модель описания речевых образцов в условиях проявления разных видов эмоций. Модель ЭОР должна отражать взаимосвязь вида эмоций и объективных характеристик РС. На настоящий момент определение такой взаимосвязи вызывает затруднение у большинства исследователей в этой области.

В данной статье делается попытка систематизации и анализа объективных признаков ЭОР. Основная задача получения признаков ЭОР состоит в том, чтобы преобразовать звуковую волну в такое признаковое пространство, в котором множество объектов одного класса будет сгруппировано вместе, а множество объектов альтернативных классов максимально разнесено. Соотнесение распознаваемого объекта (в

данном контексте под объектом понимается фонема РС [1]) с базой объектов, которые необходимо идентифицировать, проходит в три этапа: 1) выделение того или иного признака объекта; 2) объединение признаков в комплексы или классы; 3) выбор предполагаемого значения из ряда альтернатив. Литературный обзор, охватывающий результаты исследований отечественных и зарубежных авторов [1 − 11] показывает, что на данном этапе можно выделить четыре группы объективных признаков, позволяющих различать речевые образцы: спектрально-временные, кепстральные, амплитудно-частотные и признаки нелинейной динамики (табл.). Рассмотрим подробно каждую группу признаков.

Таблица. Признаки ЭОР

Название признака	Обозначение	Область		Исследования
		Синтез	Распознавание	
1	2	3	4	5
I. Спектрально-временные признаки				
I.I. Спектральные признаки				
1) Среднее значение спектра анализируемого речевого сигнала	$X(i)$	+	+	[3]
2) Нормализованные средние значения спектра	$X_H(i)$	+	+	
3) Относительное время пребывания сигнала в полосах спектра	$t(i)$	+	+	
4) Нормализованное время пребывания сигнала в полосах спектра	$t_H(i)$	+	+	
5) Медианное значение спектра речи в полосах	$m_H(i)$	+	+	
6) Относительная мощность спектра речи в полосах	$P_H(i)$	+	+	
7) Вариация огибающих спектра речи	$V(i)$	+	+	
8) Нормализованные величины вариации огибающих спектра речи	$V_H(i)$	+	+	
9) Коэффициенты кросскорреляции спектральных огибающих между полосами спектра	$R(i,k)$	+	+	
I.II. Временные признаки				
10) Длительность сегмента, фонемы	l	+	+	[1, 3, 5, 7, 9, 11]
11) Высота сегмента	h	+	+	[1]
12) Коэффициент формы сегмента	k	+	+	
II. Кепстральные признаки				
13) Мелко частотные кепстральные коэффициенты	*MFCC*	−	+	[6, 8]
14) Коэффициенты линейного предсказания с коррекцией на неравномерность чувствительности человеческого уха	*PLP*	−	+	

Продолжение таблица

1	2	3	4	5
15) Коэффициенты мощности частоты регистрации	*LFPC*	−	+	[6]

16) Коэффициенты спектра линейного предсказания	LPC	–	+	
17) Коэффициенты кепстра линейного предсказания	$LPCC$	–	+	
III. Амплитудно-частотные признаки				
18) Интенсивность, амплитуда	i, A	–	+	[5, 8, 9]
19) Энергия	E	+	+	[7, 11]
20) Частота основного тона (ЧОТ)	$F0$	+	+	[3, 4, 6, 7, 8, 10, 11]
21) Формантные частоты	$F1, F2, F3, F4$	+	+	[3, 5, 6, 11]
22) Джиттер	Ji	–	+	[4, 6]
23) Шиммер	Sh	–	+	[4, 6, 8]
24) Радиальная базисная ядерная функция	$K(x, y)$	–	+	[10]
25) Нелинейный оператор Тигера	TEO	–	+	[4, 8, 10]
IV. Признаки нелинейной динамики				
26) Отображение Пуанкаре	Δt_i	–	+	[2]
27) Рекуррентный график	$R_{i,j}$	–	+	
28) Максимальный характеристический показатель Ляпунова	Yj	–	+	[2, 5]
29) Фазовый портрет (аттрактор)	Y_n	–	+	
30) Размерность Каплана-Йорка	D	–	+	[5]

Спектрально-временные признаки [3] характеризуют РС в его физико-математической сущности исходя из наличия компонентов трех видов:

1) периодических (тональных) участков звуковой волны; 2) непериодических участков звуковой волны (шумовых, взрывных); 3) участков, не содержащих РС (речевых пауз). Спектрально-временные признаки позволяют отражать своеобразие формы временного ряда и спектра голосовых импульсов у разных лиц и особенности фильтрующих функций их речевых трактов. Характеризуют особенности речевого потока, связанные с динамикой перестройки артикуляционных органов речи говорящего, и являются интегральными характеристиками речевого потока, отражающими своеобразие взаимосвязи или синхронности движения артикуляционных органов говорящего.

В группе спектрально-временных признаков РС рассматривается как некоторый квазистационарный процесс [1]. Среди множества акустических параметров были выделены параметры, инвариантные к действию повышенного уровня сигнала, описывающие статистические характеристики РС и основного тона, особенности спектральной структуры. РС представлен в виде последовательности значений кратковременных энергетических спектров, измеренных в моменты времени $j = 0,1,..., J$ каждые 5,7 мс (значение 5,7 мс выбрано экспериментально). РС подвергается спектральному анализу посредством быстрого преобразования Фурье (БПФ). С помощью БПФ спектры вычисляются последовательно по РС с применением набора 24 фильтров, соответствующим 24 критическим полосам. Таким образом, РС представляется в виде

$$\{x(0, j),...x(i, j),...x(23, j)\}, j = 0,1,..., J,$$

где $x(i, j)$ – значение сигнала на выходе i-го полосового фильтра в j-м кратковременном энергетическом спектре; J – общее количество спектральных срезов на анализируемом отрезке.

В ходе проведенных нами исследований по группе спектрально-временных признаков была экспериментально доказана взаимосвязь временных параметров модели, а именно длительности сегментов l, их отношений dl и средних значений длительности сегментов $\bar{l}$ со степенью эмоционального окраса РС. Это позволяет использовать временные характеристики РС в качестве признаков, позволяющих распознавать ЭОР. Доказано наличие взаимосвязи между субъективной (воспринимаемой слушателями) качественной оценкой образцов речи и объективными количественными параметрами РС.

Установлено, что применение только спектральных характеристик невозможно использовать в качестве признаков, позволяющих правильно распознавать и идентифицировать различные эмоции [12].

Следует особо отметить такую группу признаков ЭОР, как кепстральные коэффициенты [6, 8]. Большинство современных автоматических систем синтеза и распознавания речи сосредотачивают усилия на извлечении частотной характеристики речевого тракта человека, отбрасывая при этом характеристики сигнала возбуждения. Это объяснено тем, что коэффициенты первой модели обеспечивают лучшую разделимость звуков. Для отделения сигнала возбуждения от сигнала речевого тракта прибегают к кепстральному анализу. Схематически этот метод представлен на рисунке.

Общая схема кепстрального анализа сигнала: FFT − блок быстрого преобразования Фурье сигнала (БПФ); LOG я блок логарифмирования спектра; IFFT − блок обратного быстрого преобразования Фурье (ОБПФ)

Линейное предсказание является одним из наиболее эффективных методов анализа РС. Этот метод является доминирующим при оценке основных параметров РС, таких как, например, период основного тона, форманты, спектр, функция площади речевого тракта, а также при сокращенном представлении речи с целью ее передачи и хранения. Важность метода обусловлена высокой точностью получаемых оценок и простотой вычислений. Основной принцип метода линейного предсказания состоит в том, что текущий отсчет РС можно аппроксимировать линейной комбинацией предшествующих отсчетов. Коэффициенты предсказания при этом определяются однозначно минимизацией среднего квадрата разности между отсчетами РС и их предсказанными значениями.

Мел-частотные кепстральные коэффициенты *MFCC* широко используются в качестве набора признаков РС. Основной идеей метода *MFCC* является максимальное приближение информации поступающей на слуховой анализатор мозга человека. Признаки, построенные на основе *MFCC*, учитывают психоакустические принципы восприятия речи, поскольку используют мел-шкалу, связанную с критическими полосами слуха и вычисляются следующим образом:

$$MFCC = \sum_{l=1}^{L} E_l \cos(k(l-0{,}5)\frac{\pi}{l}),\ 1 \le k \le L,$$

$$E_l = \lg\left(\sum (S(k))^2 \omega(k - (k1_l + \frac{\Delta K_l}{2})) \right),$$

где $S(k)$ − спектр Фурье; $k1_l, k2_l$ − границы частотных диапазонов l-й мелко частотной полосы; $\Delta K_l = k2_l - k1_l$ − четное число; $\omega(x)$ − оконная функция; L − количество мел-полос.

Все перечисленные кепстральные коэффициенты позволяют уменьшить размерность исходного признакового пространства, что скажется на быстродействии вычисления различных параметров ЭОР.

Экспериментальные исследования по кепстральным признакам нами не проводились.

Анализ амплитудно-частотных параметров РС позволяет применять эти характеристики в качестве информативных признаков диагностики ЭС человека и синтеза ЭОР [3 − 11]. Амплитудно-частотные признаки позволяют получать оценки, значения которых могут меняться в зависимости от параметров дискретного преобразования Фурье (вида и ширины окна), а также при незначительных сдвигах окна по выборке. РС акустически представляют собой распространяемые в воздушной среде сложные по своей структуре звуковые колебания, которые характеризуются в отношении их частоты (числа колебаний в секунду), интенсивности (амплитуды колебаний) и длительности. Все эти характеристики подвержены изменениям на протяжении одного РС. Они могут быть зафиксированы и измерены посредством специальных электронно-акустических приборов (прежде всего осциллографа и спектрографа). Амплитудно-частотные признаки несут необходимую и достаточную информацию для человека по РС при минимальном времени восприятия.

Несмотря на многочисленные работы в данной области, специальных исследований на взаимосвязь эмоций с амплитудно-частотными параметрами пока практически не проводилось. Проведенные нами исследования позволили выявить тот факт, что применение этих признаков не позволяет в полной мере использовать их в качестве инструмента идентификации ЭОР [12].

Для группы признаков нелинейной динамики [2, 5] РС рассматривается как скалярная величина, наблюдаемая в системе голосового тракта человека. Нелинейные процессы играют важную роль в речеобразовании, процесс речеобразования можно считать нелинейным и, следовательно, анализировать его методами нелинейной динамики.

Задача нелинейной динамики состоит в нахождении и подробном исследовании базовых математических моделей и реальных систем, которые исходят из наиболее типичных предложений о свойствах отдельных элементов, составляющих систему, и законах взаимодействия между ними.

В настоящее время методы нелинейной динамики базируются на фундаментальной математической теории, в основе которой лежит теорема Такенса [5], которая подводит строгую математическую основу под идеи нелинейной авторегрессии и доказывает возможность восстановления фазового портрета аттрактора по временному ряду или по одной его координате. Под аттрактором понимают множество точек или подпространство в фазовом пространстве, к которому приближается фазовая траектория после затухания переходных процессов. Оценки характеристик сигнала из восстановленных речевых траекторий используются в построении нелинейных детерминированных фазово-пространственных моделей наблюдаемого временного ряда.

Экспериментально подтверждено, что выявленные отличия в форме аттракторов Y_n можно использовать для диагностических правил и признаков, позволяющих распознать и правильно идентифицировать различные эмоции в эмоционально окрашенном РС. Также выявлено, что рекуррентный график $R_{i,j}$ можно использовать в качестве признака, позволяющего правильно распознавать эмоции по РС, распознавание эмоций проводится по характеру распределения точек в квадрате рекуррентного графика $R_{i,j}$.

Таким образом, выполнен анализ объективных признаков, используемых для распознавания эмоций в речи. Проведены вычислительные эксперименты по исследованию эмоций в РС, с целью выявления и отбора наиболее информативных признаков. В результате проведенного исследования подготовлена платформа объективных признаков РС для синтеза ЭОР и разработки методов распознавания ЭС человека.

Библиографический список

1. Калюжный, М.В. Система реабилитации слабовидящих на основе настраиваемой сегментарной модели синтезируемой речи: дис. …канд. тех. наук / М.В. Калюжный. СПб., 2009.
2. Перервенко, Ю.С. Исследование инвариантов нелинейной динамики речи и принципы построения системы аудиоанализа психофизиологического состояния: дис. …канд. тех. наук / Ю.С. Перервенко. Таганрог, 2009.
3. Розалиев, В.Л. Моделирование эмоциональных реакций пользователя при речевом взаимодействии с автоматизированной системой: дис. …канд. тех. наук / В.Л. Розалиев. Волгоград: ВГТУ, 2009.
4. Соловьева, Е.С. Методы и алгоритмы обработки, анализа речевого сигнала для решения задач голосовой биометрии: дис. …канд. тех. наук / Е.С. Соловьева. М., 2008.
5. Хроматиди, А.Ф. Исследование психофизиологического состояния человека на основе эмоциональных признаков речи: дис. …канд. тех. наук / А.Ф. Хроматиди. Таганрог, 2005.
6. Chen, Y.T. A study of emotion recognition on mandarin speech and its performance evaluation: Ph. D. dissertation / Y.T. Chen. Tatung, 2008.
7. Morist, M.U. Emotional speech synthesis for a radio dj: corpus design and expression modeling: master thesis MTG-UPF dissertation / M.U. Morist. Barcelona, 2010.
8. Siging, W. Recognition of human emotion in speech using modulation spectral features and support vector machines: master of science dissertation / W. Siging. Kingston, 2009.
9. Алдошина, И.А. Связь аккустических параметров с эмоциональной выразительностью речи и пения / И.А. Алдошина // Звукорежиссер. № 2. СПб., 2003.
10. Хейдоров, И.Э. Классификация эмоционально окрашенной речи с использованием метода опорных векторов / И.Э. Хейдоров, Я. Цзинбинь, [и др.] // Речевые технологии. Вып. 3. СПб., 2008. С. 63−71.
11. Makarova, V. RUSLANA: a database of russian emotional utterances / V. Makarova, V.A. Petrushin // ICSLP, 2002. P. 2041−2044.
12. Сидоров, К.В. К вопросу оценки эмоциональности естественной и синтезированной речи по объективным признакам / К.В. Сидоров, М.В. Калюжный // Вестник Тверского государственного технического университета. Вып. 18. Тверь, 2011. С. 81−85.