А.В. Вартанов

A.V.Vartanov

кандидат психологических наук, старший научный сотрудник
МГУ имени М.В.Ломоносова, факультет психологии, кафедра психофизиологии
Lomonosov Moscow State University, Faculty of Psychology, Department of Psychophysiology

Антропоморфный метод распознавания эмоций в звучащей речи[1]
Anthropomorphic method of emotion recognition in sounding speech

Аннотация

Предложен новый эффективный метод автоматического распознавания эмоций по речевому сигналу, основанный на четырехмерной сферической модели эмоций и принципах кодирования информации в нервной системе. В результате разработан и экспериментально протестирован принцип относительного кросс частотного амплитудно-вариабельного кодирования эмоций в речевом сигнале. Проверялась гипотеза о речи как о многоканальном (разнесенном по частотам) сигнале, в каждой полосе которого возможны независимые быстрые микро изменение амплитуды. Показано хорошее согласие выделенных параметров речевого сигнала и субъективного восприятия тех же образцов (коротких слов «да» и «нет») в системе формализованных параметров четырехмерной психофизиологической модели эмоций. Полученные параметры (факторы) можно охарактеризовать как бимодальные спектральные фильтры. Фактор 1 имеет основной максиму в области 3000 Гц и вспомогательный – на 500 Гц. Он определяет изменение звукового сигнала по оси знака эмоций – чем более вклад данного компонент по сравнению с другими компонентами, тем положительнее (лучше, полезнее) оценивается объект высказывания. Фактор 2 имеет два близких максимума на частотах 1000 и 1750 Гц. Он определяет степень информационной неопределенности ситуации – удивление в противоположность уверенности (спокойствию). Фактор 3 имеет самые широко разнесенные максимумы – в низкочастотной области (около 150 Гц) и высокочастотной области (3500 Гц). Он характеризует притяжение (любовь), при этом для набора слов «нет» он сопровождается отсутствием активного отвержения, а для набора «да» – положительной оценкой (знаком). Фактор 4 имеет близкие максимумы на 600 и 1500 Гц, по конфигурации близок к фактору 2, но сдвинут относительно него в низкочастотную область, попадая своими максимумами в его локальные минимумы. Этот компонент соответствует характеру отвержения, определяет, будет ли агрессивная (активная) или пассивная (страх, бегство) реакция. Полученные результаты в целом подтверждают продуктивность предлагаемого антропоморфного подхода к разработке технических систем, в частности методам обработки речевого сигнала и представления данных. Обнаруженное совпадение подтверждает и выделенные ранее параметры психофизиологической модели, дополнительно обосновывая предпочтительность (по сравнению с другими известными в

[1] Работа поддержана РФФИ проект № 11-06-12036-офи-м-2011.

литературе) именно такой системы классификации эмоций, как с точки зрения размерности, так и в отношении ориентации осей пространства модели.

Abstract

A new efficient method for automatic recognition of emotion in speech signal, based on the four-dimensional spherical model of emotions and principles of information coding in the nervous system. As a result, proposed and experimentally tested the principle of the relative cross-frequency amplitude-variable encoding of emotions in speech signal. Testable hypotheses about the speech as multichannel (frequency diversity) signal in each band is possible independent fast micro amplitude change. Good agreement between the selected parameters of the speech signal and the subjective perception of the system parameters formalized four dimensional model of psychophysiological emotion of the same samples (short words "yes" and "no"). The obtained parameters (factors) may be characterized as bimodal spectral filters. Factor 1 has a basic maxim in the 3000 Hz and a secondary – 500 Hz. It determines the change in the sound signal in accordance with an axis «character emotion» – than the contribution of this component as compared with other components, the more positive (better useful) is estimated to utterance. Factor 2 has two maxima at frequencies near 1000 and 1750 Hz. It determines the degree of information uncertainty of the situation – as opposed to the surprise of confidence (calm). Factor 3 is the most widely spaced peaks – at low frequencies (about 150 Hz) and high-frequency region (3500 Hz). He characterizes the attraction (love). In this set of words for "no" it is accompanied by the absence of active rejection, and for a set of "yes" – a positive assessment. Factor 4 has similar maxima at 600 and 1500 Hz. Configuration it is close to factor 2, but it is shifted with respect to the low-frequency region, getting their peaks in its local minima. This component corresponds to the nature of rejection determines whether aggressive (active) or passive (fear, escape) reaction. The results obtained confirm the efficiency of the proposed general anthropomorphic approach to the development of technical systems, in particular the methods of speech signal processing and data presentation. And confirms the identity of a previously identified psychophysiological model parameters, further justifying the preference (compared with other well known in the literature) is a system of classification of emotions, both in terms of dimensions and orientation of the axes with respect to the model space.

Ключевые слова: эмоции, речевой сигнал, антропоморфный метод

Keywords: emotions, speech signal, the anthropomorphic approach

Введение в проблему

Давно известно, что речь, порождаемая человеком, когда он находится в различных эмоциональных состояниях, различается по целому ряду показателей. К числу таких наиболее информативных характеристик относят, прежде всего, характеристики просодической

группы, которые тонко отражают процессуальную сторону устных высказываний и в первую очередь изменяются при реакциях аффективного плана (Златоустова, 1957; Михайлов, Златоустова, 1987; Никишкян, 1987). При этом задача автоматического распознавания звучащей речи и, в частности, ее эмоциональной окрашенности является междисциплинарной и постоянно привлекает исследователей разных специальностей – не только лингвистов, но и математиков, программистов, психологов, физиологов. От ее решения зависит прогресс современных автоматизированных систем управления, реабилитации и протезирования, систем безопасности, срочного оповещения и т.п. Решение этой задачи имеет большое научное значение во всех сферах фундаментальных исследований человека и информационных технологий. При этом, как отмечается в (Сидоров, Филатова, 2012), в последние годы явно усилился интерес к анализу речевого сигнала, рассматриваемого в качестве наиболее удобного объективного показателя выражения эмоций, эмоционального состояния человека. Это касается не только сфер с повышенной ответственностью деятельности – космонавтика, авиация (летчики, диспетчера аэропорта), обслуживание АЭС, и т.д. (Хроматиди, 2005; Соловьева, 2008; Chen, 2008; Siging, 2009; Фролов, Милованова, 2009; Розалиев, 2009; Калюжный, 2009; Перервенко, 2009; Morist, 2010), которые изначально доминировали в этом отношении, но и в широкой бытовой сфере. В интернете, новостных лентах и других популярных изданиях периодически появляются сообщения о все более успешных попытках создания программ и бытовых устройств, реагирующих на эмоции в голосе человека. Например, «Ноосфера» (Шпикуляк, 2012) сообщает, что «инженеры из Рочестерского университета (Великобритания) разработали программу, способную распознавать эмоции человека по его речи, даже не понимая смысла сказанного. Программа ориентируется на базу звукозаписей, состоящую из календарных дат, произнесенных профессиональными актерами с разными интонациями. Алгоритм анализирует 12 характерных параметров речи, таких как высота и громкость звука. На их основании он определяет одну из шести эмоций. По словам разработчиков, точность распознавания составляет 81 процент — значительно лучше 55 процентов, которых удавалось добиться в предыдущих аналогичных исследованиях. Авторы уже разработали первое коммерческое приложение — программу, отображающую на экране веселый или грустный смайлик в зависимости от результата анализа записанного голоса. Это лишь первый этап. Авторы программы фантазируют, что в дальнейшем смартфоны смогут менять цветовую схему интерфейса или выбирать подходящую музыку в зависимости от настроения владельца». На сайте (Animal language, 2013) отмечается, что хотя изучение языка эмоций точными научными методами еще лишь начинается, но уже сейчас вырисовываться большое значение этой проблемы, как для теоретической науки, так и для практики. При этом уже всем ясно, что решить эту задачу нельзя без знания алфавита акустического языка эмоций. Но «чтобы

заложить этот алфавит в электронный мозг робота, необходимо формализовать признаки, ответственные за эмоциональность голоса» (Animal language, 2013). Однако, как отмечается в (Сидоров, Филатова, 2012), несмотря на множество исследований и коммерческих предложений в данной области, проблема автоматического распознавания эмоционального состояния говорящего по речи на данный момент не является полностью решенной, в частности, отсутствует модель описания речевых образцов в условиях проявления разных видов эмоций. Процесс интерпретации (распознавания) эмоций человека по естественной речи является весьма сложной задачей, как в плане математической формализации задачи, так и в способах четкой конкретизации эмоционального состояния – однозначного детектирования эмоции по речевому сигналу (Филатова, Сидоров, 2012). И, как отмечают эти авторы, в настоящее время отсутствует универсальная теоретическая модель описания речевых образцов в условиях проявления разных видов эмоций.

Это обусловлено целым комплексом взаимосвязанных проблем. С одной стороны необходимо выделить в речевом сигнале те параметры, которые могли бы служить индикаторами эмоций – это проблемы регистрации, математического анализа и соответствующих алгоритмов и технических средств. Но для решения этой задачи требуется четко задать «входные» и «выходные» данные, формально представить требуемый результат. А с другой стороны, необходимы формальные, объективные методы для систематизации и классификации таких сложных явлений как эмоции человека, необходимо разработать адекватную модель и определенным образом собрать базу данных – набор соответствующих «образцов» состояний и корреспондирующих им фрагментов речи. Получается порочный круг: чтобы решить одну задачу надо уже иметь решение другой. Тем не менее, научные исследования и практические разработки в этом направлении предпринимаются все с большей интенсивностью, подстегиваемой коммерческими возможностями. При этом, как правило, разработчики новых методов и инструментов анализа пользуются лишь собственным «здравым смыслом» и некоторыми теоретическими обобщениями психологов и фонологов. А последним для анализа эмоциональных явлений приходится пользоваться «стандартными», общедоступными инструментами объективного анализа речевых сигналов. И чтобы хоть как-то приблизиться к практической эффективности всем приходится упрощать задачу – при разработке новых методов анализа речевого сигнала ограничиваться отдельными аспектами эмоциональных феноменов, например, только интерпретацией знака эмоций (Филатова, Сидоров 2012) или отдельных эмоций, наиболее важных для данной области применения. В итоге общая эффективность предлагаемых в настоящий момент средств невысока. Приведенный выше пример из «Ноосферы» наглядно это подтверждает: даже задача ставиться по распознаванию всего шести эмоций, результат распознавания сводиться к примитивному бинарному действию. А метод в типичном случае базируется на стандартных

алгоритмах сопоставления с образцом в расчете на тупое количественное увеличение быстродействия и объема памяти (например, за счет «облачных» технологий) и размера «словаря» образцов.

О параметрах речевого сигнала

Литературный обзор, проведенный (Сидоров, Филатова, 2012) показывает, что на современном этапе можно выделить четыре группы объективных признаков и соответствующих методов, позволяющих различать речевые образцы: спектрально-временные, кепстральные, амплитудно-частотные и признаки на основе нелинейной динамики. Показано, что на основе одних только простых спектральных характеристик звукового сигнала невозможно правильно распознавать и идентифицировать различные эмоции (Сидоров, 2011). 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 артикуляторных органов говорящего. Амплитудно-частотные признаки также несут важную информацию. Как показано в (Адашинская, Чернов, 2007), большинство исследований в качестве наиболее информативных акустических коррелятов эмоциональных и функциональных состояний рассматривает ряд частотных, временных и мощностных характеристик голосового сигнала. Как правило, стенические состояния ведут к возрастанию, а астенические – к понижению показателей основного тона, формант и интенсивности. Обнаружена взаимосвязь акустических параметров речи эмоциональных и функциональных состояний, обусловленная индивидуальными особенностями говорящих, что выражается в разнонаправленности изменений ряда временных и мощностных параметров речи (Адашинская, Чернов, 2007). Однако применение этих признаков не позволяет в полной мере использовать их в качестве инструмента идентификации эмоционально окрашенной речи (Сидоров, 2011). В группе спектрально-временных признаков были выделены параметры, инвариантные к действию повышенного уровня сигнала, описывающие статистические характеристики речевого сигнала и основного тона, особенности спектральной структуры (Розалиев, 2009). Группа признаков эмоционально окрашенной речи по кепстральным коэффициентам позволяет отделить сигнал возбуждения от сигнала речевого тракта. Мел-частотные кепстральные коэффициенты широко используются в качестве набора признаков речевого сигнала, поскольку они учитывают психоакустические принципы восприятия речи и мел-шкалу, связанную с критическими полосами слуха (Siging, 2010; Сидоров, Филатова, 2012). Для группы признаков нелинейной динамики (Старченко и др., 2010) речевой сигнал рассматривается как скалярная величина,

наблюдаемая в системе голосового тракта человека. В настоящее время методы нелинейной динамики и нелинейной авторегрессии позволяют восстанавливать фазовый портрет аттрактора по временному ряду или по одной его координате. Экспериментально подтверждено, что выявленные отличия в форме аттракторов можно использовать для диагностических правил и признаков, позволяющих распознать и правильно идентифицировать различные эмоции в эмоционально окрашенном речевом сигнале. Так, в работе (Филатова, Сидоров, 2012) предложена модель интерпретации знака эмоции по правилу объединения нечетких множеств, характеризующих значения Rmax – усредненного максимального вектора реконструкции аттрактора по четырем квадрантам. В работе (Романенко, 2010) рассмотрена возможность применения вейвлет-анализа речевого сигнала с целью использования в системе распознавания речи. Предлагается также проводить классификацию эмоционально окрашенной речи с использованием метода опорных векторов (Хейдоров, 2008). В целом, как отмечалось уже около десяти лет назад (Бабин, Мазуренко, Холоденко, 2004) аппарат акустического анализа речи уже достаточно развит и практически все наиболее часто используемые способы расчета акустических параметров речевого сигнала реализованы в известных и общедоступных математических компьютерных пакетах обработки сигналов, например, в пакетах SPL и IPPS фирмы Intel (Intel Developer Centers, 2013).

Таким образом, речь, порождаемая человеком, находящимся в различных эмоциональных состояниях, характеризуется целым рядом показателей, в том числе таких, которые могут отражать процессуальную сторону устных высказываний. Однако, формальные критерии, хотя и позволяющие успешно дифференцировать отдельные эмоции по речевым образцам, не могут дать общей картины изменения текущего состояния и отношения человека, поскольку не разработана антропоморфная система классификации эмоциональных проявлений в звучащей речи. Отправной точкой решения вышеописанной проблемы должна стать система, достаточно полно моделирующая процесс восприятия эмоций человеком, которая учитывает совокупность разных аспектов их проявления, в том числе в речи. Многомерность эмоций, их проявление на различных уровнях отражения и деятельности, способность к слиянию и образованию сочетаний исключают возможность их простой линейной классификации (Вилюнас, 1984) или создания конечного дискретного набора определенных вариантов. Обычно выделяют как минимум десять типов эмоциональных отношений или так называемых фундаментальных эмоций, между которыми, однако, возможны плавные переходы. Эти типы в достаточной мере условны, обозначая (в виде понятийных категорий) лишь наиболее важные места эмоционального континуума. Поэтому, в разное время, на основе различных экспериментальных методов и эмпирических фактов делались попытки выделить в этом разнообразии ограниченное число базовых

факторов или основных "компонентов эмоционального качества", которые бы выступали по отношению к отдельным эмоциональным переживаниям как родовые исходные характеристики или «образующие». В настоящее время известен целый ряд таких независимых или частично перекрывающихся признаков и оснований для деления эмоциональных явлений. Это объясняется тем, что эмоции проявляются одновременно и во внутренних переживаниях, и в поведении, причем и то и другое обусловлено еще специфической физиологической активацией. При этом аппарат анализа речевого сигнала также должен, хотя бы в некоторой степени воспроизводить процессы, позволяющие нервной системе человека правильно распознавать всю гамму эмоций, т.е. необходима антропоморфная модель эмоций.

Четырехмерная сферическая модель эмоций

Не смотря на всю сложность проблемы, предпринятое ранее исследование эмоциональных характеристик звучащего слова и семантики эмоций позволили построить универсальную четырехмерную сферическую модель эмоций (Виденеева, Хлудова, Вартанов, 2000; Вартанов, Виденеева, 2001; Вартанов, Вартанова, 2003; Вартанов, Вартанова, 2005). Эта модель объективирует и формализует в системе четырех количественных параметров все многообразие переживания и различные проявления эмоций в речи, мимике, а также в семантике. Построение модели проводилось экспериментально с помощью многомерного шкалирования субъективных различий между эмоциональными состояниями, задаваемыми специально созданными образцами. С целью уровнять и сделать определенным содержание этих образцов, в эксперименте использовалось одно и то же слово, произнесенное в разных эмоциональных состояниях. В одной серии использовалось слово "Да", а в другой "Нет". Уже такие короткие одноударные слова, как свидетельствует практика актерского мастерства (Станиславский, 1959) вполне могут адекватно и полно отражать весь спектр эмоциональных проявлений. Эти слова по сравнению с другими несут более определенное и независящее от контекста значение, но в то же время они более нейтральны и допускают больше вариантов эмоциональной окраски в их произнесении. Из большого числа образцов, наигранных профессиональными актерами и «подловленных» в естественных условиях, были отобраны по 20 наиболее удачных в каждом наборе, отражающих 10 типичных эмоций, наиболее существенных для актерского исполнения (Станиславский, 1959). Наличие двух наборов таких образцов (противоположных по семантике) позволяет найти универсальные, не зависимые от конкретного слова параметры, определяющие именно проявление эмоций в речи. В эксперименте регистрировались субъективные оценки степени попарного различия между звуковыми стимулами. Набор из 20 образцов в каждой из серий образовывал по 190 вариантов пар. Каждая пара предъявлялась не менее чем по 3 раза, т.е. всего 570 пар, которые следовали в случайном порядке. В экспериментах участвовало в общей сложности 25

взрослых испытуемых и 30 детей разных возрастов (с 1-го по 8-й классы). Кроме того, тем же методом исследовалась и семантика эмоций русского языка, для чего использовались различные наборы слов, обозначающих эмоции. Обнаружено, что и дети и все взрослые одинаково успешно воспринимают и непосредственно сравнивают эмоциональные состояния другого, выраженного в интонациях речи – полученные матрицы всех испытуемых хорошо совпадали (коррелировали) друг с другом, что позволило далее объединить все данные и тем самым уменьшить случайный шум получаемых оценок, образующих матрицу различий.

Анализ метрическим методом многомерного шкалирования усредненных матриц различий в соответствующих сериях показал, что размерность полученного эмоционального пространства по всем критериям должна быть оценена как равная четырем. Расположение точек-стимулов в четырехмерном пространстве проверялось на сферичность. Оказалось, что в серии "Да" вариативность радиуса четырехмерной сферы составляла всего 9.71%, а в серии "Нет" - 9.94%. Это хорошо согласуется с теоретическими разработками о принципах кодирования в нервной системе (Соколов, Вайткявичюс, 1989; Соколов, 2001; Вартанов, 2011), на основе которых может быть построена антропоморфная и нейротропная модель эмоций.

После вращения, евклидовы оси пространства получили интерпретацию как определенные нейронные (мозговые) механизмы эмоций, а угловые характеристики – как субъективные качества эмоций. Первые две евклидовы оси пространства модели связаны с оценкой ситуации: ось 1 – по знаку (хорошо, полезно, приятно или плохо, вредно, неприятно), ось 2 – по степени информационной определенности (уверенность – удивление). Система третьей и четвертой осей связана с побуждением: ось 3 – притяжение, ось 4 – отвержение (оборонительная реакция) активное (агрессия) или пассивное избегание (страх, затаивание). Это хорошо согласуется с известными (Симонов, 1981; 2001) мозговыми механизмами эмоций: так, ось 3 и положительное направление оси 1 (вроде бы сходные качества) отражают работу разных групп нейронов гипоталамуса – побудительных и подкрепляющих, которые хотя и определяют, казалось бы, одни и те же положительные эмоциональные состояния, но находятся между собой в конкурентных отношениях (что проявляется в ортогональности осей модели). Ось 2 и отрицательное направление оси 1 можно связать с работой гиппокампа (активизирующегося в условиях информационной неопределенности) и фронтальной коры (дорсальной ее части), а также с системой миндалина – вентральная часть префронтальной коры. В целом префронтальная кора, являясь, как и гиппокамп «информационной» структурой мозга, ориентирует поведение на сигналы высоковероятных событий. Ось 4, которая делит активные и пассивные оборонительные реакции, по-видимому, также описывает активность медиального гипоталамуса, точнее двух его структур, стимуляция

которых вызывает оборонительные реакции нападения (положительное направление оси 4) или бегства, соответственно (отрицательное направление оси 4).

Оказалось, что три угла четырехмерной гиперсферы, выбранные в проекции осей 1-2, 3-4 и угол, образуемый движением точки между двумя этими плоскостями, задают такие субъективно переживаемые качества эмоций, как описанные еще В.Вундтом (1984) три качества: 1) эмоциональный тон (удовольствие – неудовольствие), 2) возбуждение – успокоение – угнетение, 3) напряжение – разрешение. При этом первый и второй углы упорядочивают все 10 основных эмоций по модальности: 5 эмоций, определяемых ситуацией и 5, определяемых собственной активностью. Но оказалось также, что при выборе другой системы угловых параметров – если взять три угла в системе осей 4-1, 3-2 и угол, образуемый движением точки между этими плоскостями, обнаруживается другая система классификации эмоций, описываемая при исследовании выражений лица – круговая система Х.Шлосберга (Schlosberg, 1941) и сферическая модель Ч.А.Измайлова (Измайлов, Коршунова, Соколов, 1999)), а также семантики Ч.Осгуд (Osgood, Suci, Tannenbaum, 1957): 1) эмоциональный тон или знак (упорядочивает 6 основных эмоций по модальности), 2) активность или яркость эмоций (возбуждение – покой) и 3) эмоциональная насыщенность (сила проявления эмоций).

Таким образом, эти данные показывают, что звучащая речь вполне определенно и достаточно точно выражает эмоциональное состояние говорящего, хорошо корреспондируя с другими важными для человека каналами – зрительному восприятию (по мимике и выразительным движениям), ощущениям своего собственного состояния в самонаблюдении, а также закрепленная в языковых терминах – общественный опыт обозначения эмоций в социальном канале коммуникации. Предлагаемая четырехмерная сферическая модель может служить общей классификационной системой для эмоциональных явлений, объединяя как физиологические представления о мозговых механизмах эмоциональной регуляции, так и известные психологические классификации, полученные на основе разных экспериментальных данных. Она количественно объясняет также все возможные нюансы и плавные взаимопереходы эмоций, представляя каждую конкретную эмоцию как линейную комбинацию выделенных основных психофизиологических параметров. По-видимому, у человека и животных существует специальный механизм эмоционального или чувственного отражения, необходимый для регуляции поведения и ориентировки в ситуации, работа которого может быть формально представлена в виде вышеописанной четырехмерной сферической модели. Именно наличие единого механизма во всех процессах позволяет представить все эмоциональные явления в одной и той же системе параметров. В результате данная модель, являясь антропоморфной (поскольку отражает субъективное отношение человека) и нейротропной (поскольку отражает нейронные механизмы) позволяет количественно описывать и наглядно представить изменения текущего состояния человека

или его эмоционального отношения, и может стать базисом при конструировании устройства, которое в удобной форме представляет детектируемые по звучащей речи эмоциональные состояния человека.

Результаты выявления параметров речевого сигнала в соответствии с предлагаемой антропоморфной моделью.

В качестве исходного материала для выявления параметров речевого сигнала, которые бы воспроизводили параметры вышеописанной сферической модели эмоций были использованы те же образцы звуковых фрагментов, что и в эксперименте с субъективными оценками – 20 образцов слова «да» и 20 образцов слова «нет» (средняя длительность 0.60 сек, стандартное отклонение 0.19 сек; минимальная длительность 0.3 сек, максимальная 0.98 сек; запись в полосе до 8000 Гц). После исследования возможных параметров, наиболее полно представляющих свойства данного набора образцов, было обнаружено, что наилучшим образом поставленной задаче соответствует показатель, вычисляемый по следующему алгоритму:

1) Для звукового фрагмента с помощью стандартных средств – быстрое преобразование Фурье со сглаживанием в минимальном скользящем окне порядка 10-15 мс вычисляется последовательность мгновенных спектров мощности сигнала (в диапазон от 0 до 4000 Гц с шагом 50 Гц).
2) На основе последовательности мгновенных спектров в скользящем окне (исследовались окна порядка 50-200 мс) вычисляется показатель микро вариативности (стандартное отклонение) амплитуды (квадратного корня от мощности) на каждой частоте.
3) Для вычисления интегральной оценки всего звукового образца использовалось простое усреднение предыдущего показателя по всему интервалу звучания и получения одного вектора (по частоте) для каждого звукового образца.

Такой алгоритм был выбран на основе теоретических предположений об общих принципах кодирования информации в нервной системе (Вартанов, 2011). Дополнительным основанием к этому послужили наблюдения, впервые сделанные еще Ч.Дарвином (Дарвин, 1940) о том, что эмоциональную выразительность голосу придает именно определенное «дрожание» тембра, что особенно важно для выразительности пения. Изменения громкости речи в «макро» варианте, на протяжении всего высказывания, как отмечалось многими авторами, также может характеризовать эмоциональное отношение говорящего. Однако и быстрые «микро» изменения амплитуды (в пределах короткого слова или междометия) также могут служить мерой изменения эмоционального состояния или отношения человека. При этом, для того, чтобы возможно было передать всю гамму эмоций, как показано выше, недостаточно только одного параметра, поэтому проверялась гипотеза о речи как о многоканальном (разнесенном по частотам) сигнале, в каждой полосе которого возможны

независимые быстрые микро изменение амплитуды. То есть, основное предложение свелось к проверке относительного кросс частотного амплитудно-вариабельного кодирования эмоций в речевом сигнале.

Все полученные звуковые образцы (40 записей разной длины) были обработаны с помощью специально созданных программных средств, а усредненные значения предлагаемого параметра в исследованном частотном диапазоне (с шагом 50 Гц) были собраны в единый массив данных, который далее подвергся статистическому (факторному) анализу. Вращение и интерпретация полученных факторов проводилась с помощью специально разработанных средств на основе сопоставления с известными для данных образцов (наборов слов «да» и «нет») оценками в четырехмерной модели эмоций.

В результате факторный анализ позволил оценить размерность и выявить 4 фактора (рис. 1), которые совокупно описывают 70.15% всей дисперсии данных.

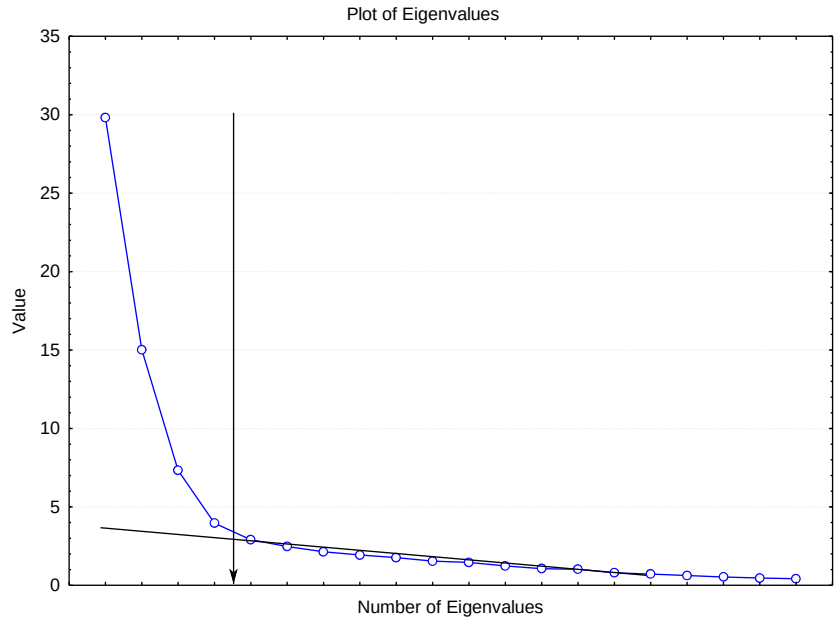


Рис. 1. График распределения собственных значений при факторном анализе всего набора звуковых образцов, включая слова «да» и «нет» (всего 40 образцов). Стрелками отмечена граница, в соответствии с которой можно оценить размерность факторного пространства как равную четырем.

После специального вращения для достижения наилучшего соответствия между нормированными значениями факторов и координатами образцов в пространстве модели эмоций, факторы получили следующее спектральное выражение, показанное на рис. 2. Решение, полученное таким методом вращения, не сильно отличалось от решения, полученного методом варимакс с нормализацией. В результате полученные факторы можно охарактеризовать как бимодальные спектральные фильтры. Фактор 1 имеет основной максиму в области 3000 Гц и вспомогательный – на 500 Гц. Фактор 2 имеет два близких максимума на частотах 1000 и 1750 Гц. Фактор 3 имеет самые широко разнесенные

максимумы – в низкочастотной области (около 150 Гц) и высокочастотной области (3500 Гц). Фактор 4 имеет близкие максимумы на 600 и 1500 Гц и близок к фактору 2, но сдвинут относительно него в низкочастотную область, попадая своими максимумами в его локальные минимумы.

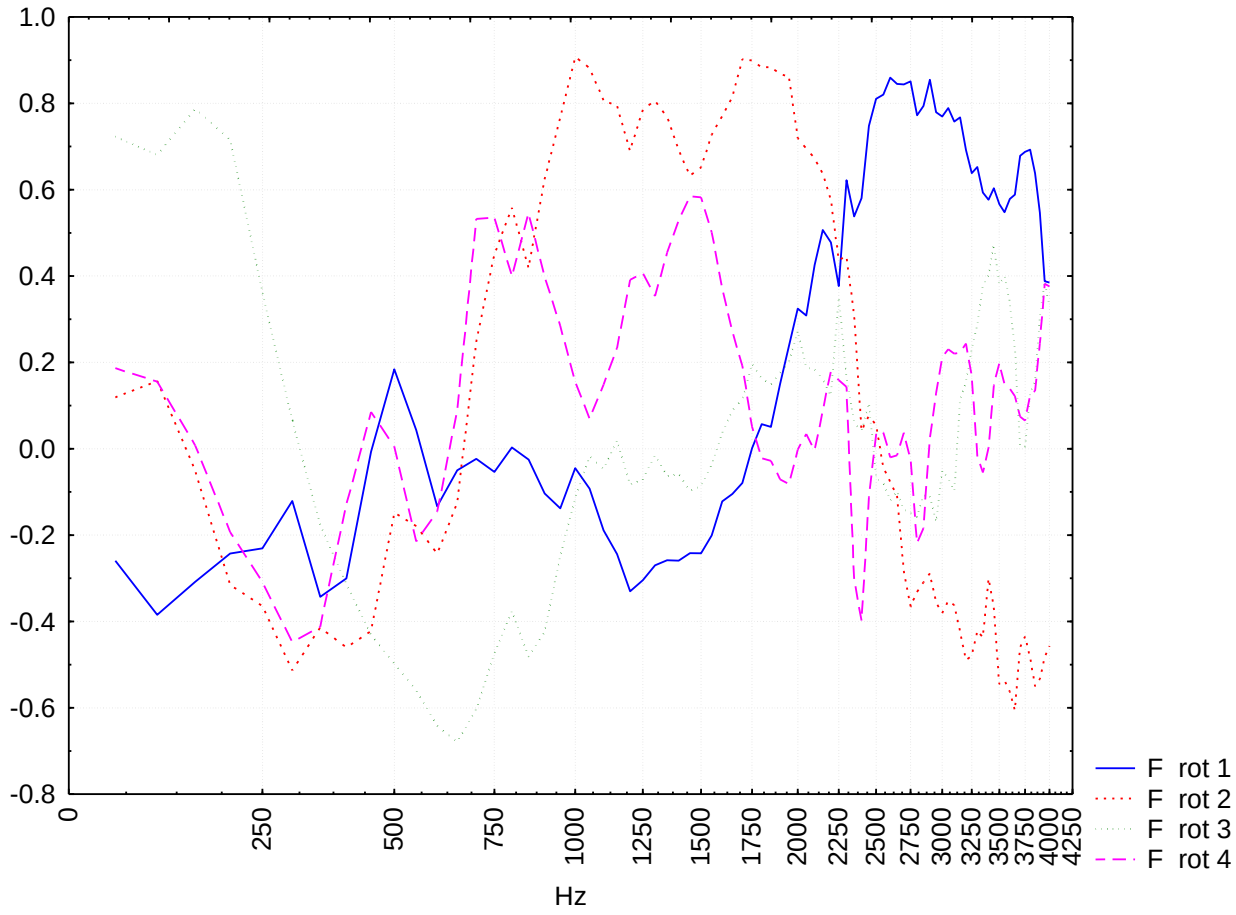


Рис. 2. Спектральные характеристики четырех выделенных факторов, которые описывают эмоциональные качества всех речевых образцов в совокупности. Горизонтальная ось дана в логарифмическом масштабе.

В результате вычисления значения этих факторов и их нормализации (как это требует теория кодирования и сферичность пространства психофизиологической модели) было проведено сопоставление оценок, полученных путем формального анализа звукового сигнала и субъективных оценок в соответствии с моделью эмоций: вычисленные коэффициенты корреляции для каждого набора в отдельности (слова «да» и «нет») и совместно представлены в таблицах 1-3.

Таблица 1. Коэффициенты корреляции Пирсона между параметрами модели эмоций (**x1-x4**) и параметрами речевого сигнала (факторами). Жирным курсивом показаны значимые ($p < .05$ при $N=40$) коэффициенты.

	x1	x2	x3	x4
Factor 1	***0.42***	***-0.36***	0.13	-0.08
Factor 2	***-0.36***	***0.59***	0.11	-0.17
Factor 3	0.11	0.11	***0.65***	-0.30
Factor 4	-0.10	-0.15	-0.27	***0.63***

Таблица 2. Коэффициенты корреляции Пирсона для набора «**да**» между параметрами модели эмоций (**x1-x4**) и параметрами речевого сигнала (факторами). Жирным курсивом показаны значимые ($p < .05$ при $N=20$) коэффициенты.

	x1	x2	x3	x4
Factor 1	***0.57***	-0.24	0.35	-0.22
Factor 2	-0.14	0.28	***0.78***	***-0.51***
Factor 3	***0.53***	0.07	***0.50***	0.04
Factor 4	-0.03	-0.41	***-0.57***	***0.72***

Таблица 3. Коэффициенты корреляции Пирсона для набора «**нет**» между параметрами модели эмоций (**x1-x4**) и параметрами речевого сигнала (факторами). Жирным курсивом показаны значимые ($p < .05$ при $N=20$) коэффициенты.

	x1	x2	x3	x4
Factor 1	0.15	***-0.46***	-0.23	0.15
Factor 2	***-0.50***	***0.55***	-0.14	-0.43
Factor 3	-0.31	0.41	***0.79***	***-0.58***
Factor 4	-0.05	-0.22	0.12	***0.55***

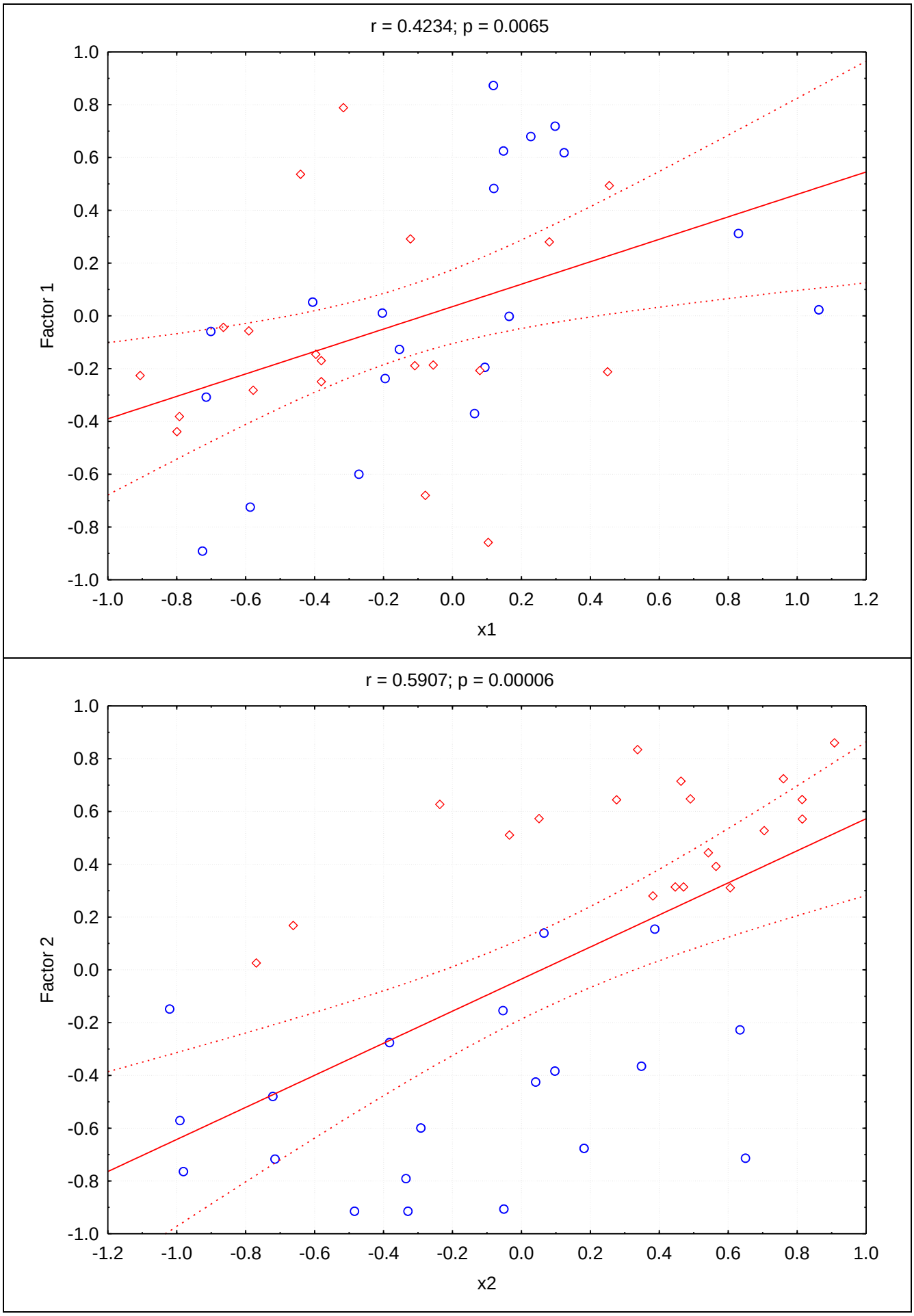
При анализе этих данных необходимо учесть, что хотя и выделенные факторы и параметры (оси) модели эмоций ортогональны, тем не менее, исследуемые образцы не заполняют все пространство равномерно и существенно различаются для наборов «да» и «нет». Поэтому сами координаты образцов в модели в некоторой степени коррелируют между собой (таблица 4). Похожая корреляция по той же причине наблюдается и между значениями факторов, что понятно, если система голосовых параметров (выделенных факторов) и система психофизиологических параметров модели близки.

Таблица 4. Коэффициенты корреляции Пирсона между параметрами модели эмоций (x1-x4). Жирным курсивом показаны значимые ($p < .05$ при $N=40$) коэффициенты.

	x1	x2	x3	x4
x1	1.00	-0.22	0.07	-0.03
x2	-0.22	1.00	***0.34***	-0.30
x3	0.07	***0.34***	1.00	***-0.44***
x4	-0.03	-0.30	***-0.44***	1.00

Корреляционные поля для выделенных факторов представлены на рис. 3. На основании этих данных можно заключить, что в целом первые четыре спектральных

параметра хорошо соответствуют (значимо коррелируют) с параметрами психофизиологической модели. При этом наблюдается определенное своеобразие связей в зависимости от набора образцов, из чего можно предположить, что семантическое значение слова («да» или «нет») в некоторой степени определяют и направление изменения данных параметров голоса. Тем не менее, можно заключить, что первый фактор определяет изменение звукового сигнала по оси знака эмоций – чем более вклад данного компонент по сравнению с другими компонентами, тем положительнее (лучше, полезнее) оценивается объект высказывания. Это более справедливо для утверждений (набор «да»). Второй спектральный параметр в целом и в наборе «нет» определяет степень информационной неопределенности ситуации – удивление в противоположность уверенности (спокойствию). При этом для слов «да» это удивление сопровождается также еще и влечением, и «не отвержением», т.е. характеризует любопытство в случае согласия или чистое удивление в случае отрицания. Третий компонент в целом и во всех наборов в отдельности характеризует притяжение (любовь), при этом для набора «нет» он сопровождается отсутствием активного отвержения, а для набора «да» положительной оценкой (знаком). Четвертый компонент соответствует как в целом, так и для обоих наборов по отдельности степени и характеру отвержения, определяет, будет ли агрессивная (активная) или пассивная (страх, бегство) реакция. При этом в наборе «да» он характеризуется еще «не притяжением».


r = 0.4234; p = 0.0065
Factor 1
x1
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
r = 0.5907; p = 0.00006
Factor 2
x2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0

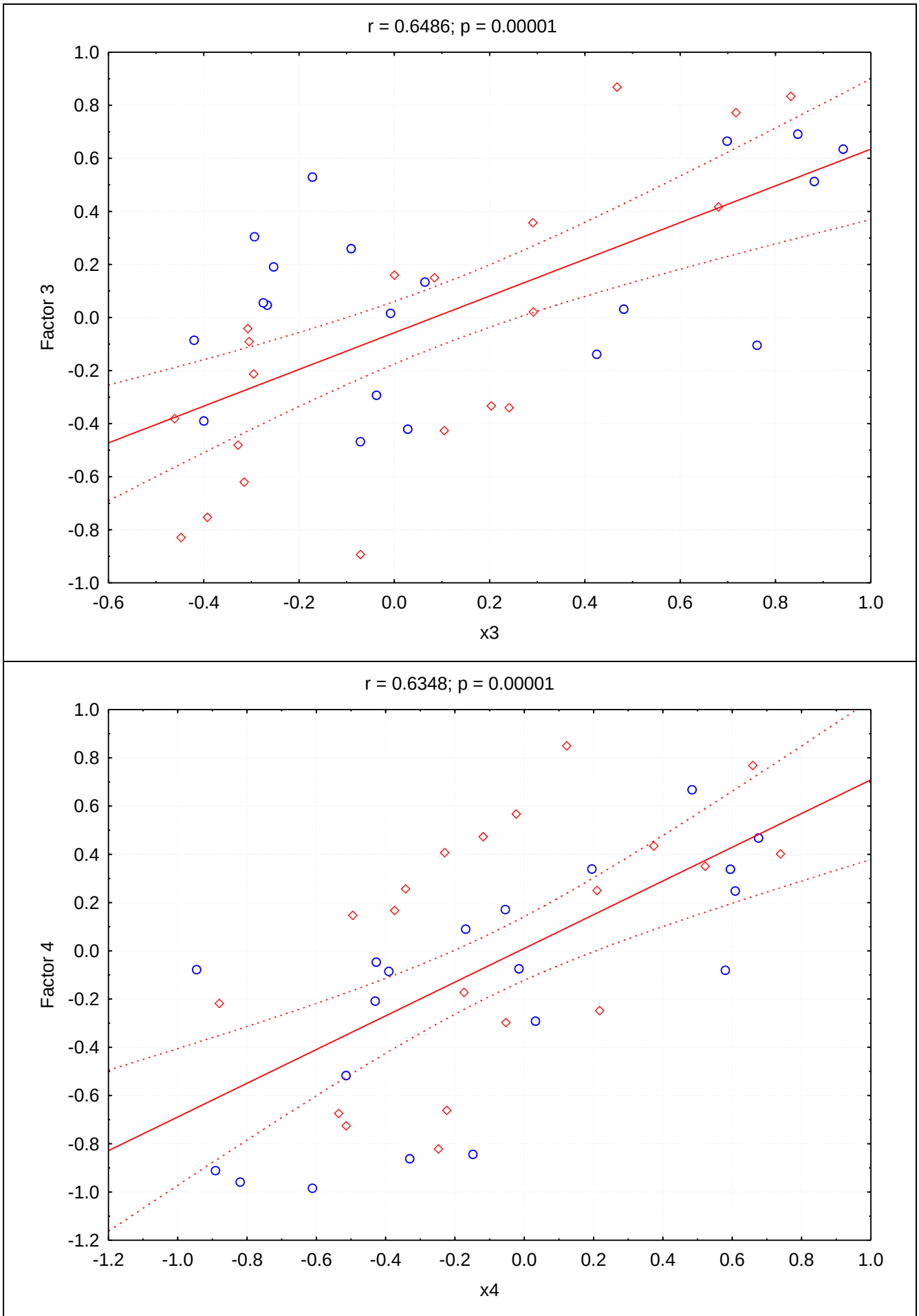


Рис. 3. Корреляционные поля соответствия выделенных параметров речевого сигнала и психофизиологической модели. Кружками обозначены образцы «да», а квадратиками – «нет».

Заключение

Проведенный анализ и полученные в результате параметры звукового сигнала позволяют построить эффективный антропоморфный (и по процессу, и по результату) метод диагностики и представления эмоций в звучащей речи. Хорошее согласие параметров речевого сигнала и психофизиологической модели эмоций подтверждают теоретические представления о принципах кодирования информации в нервной системе и продуктивности предлагаемого антропоморфного подхода к разработке технических систем, в частности методам обработки речевого сигнала. С другой стороны, обнаруженное совпадение подтверждает и выделенные ранее параметры психофизиологической модели, дополнительно обосновывая предпочтительность (по сравнению с другими известными в литературе) именно такой системы классификации эмоций, как с точки зрения размерности, так и в отношении ориентации осей пространства модели. Полученные данные также ставят новые вопросы о взаимосвязи и взаимодействии разноуровневых систем управления – вербальной сознательной и эмоциональной, досознательной, которые совместно отражаются в речевом сигнале. Необходимо также провести дополнительное исследование универсальности выделенных параметров речевого сигнала по отношению к специфике голоса диктора (в данном исследовании описан голос только одного диктора) и различные речевые высказывания, поскольку возможна интерференция содержания и просодического оформления высказывания, а также интерференция параметров, кодирующих эмоциональное и вербальное содержание речевого сигнала.

Литература.

1. Адашинская Г.А., Чернов Д.Н. Акустические корреляты индивидуальных особенностей функциональных и эмоциональных состояний. // Авиакосмическая и экологическая медицина. Т. 41 № 2. 2007, с. 3-13
2. Бабин Д.Н., Мазуренко И.Л., Холоденко А.Б. О перспективах создания системы автоматического распознавания слитной устной русской речи. // Интеллектуальные системы, том 8, выпуск 1-4, М. 2004, с. 45-70
3. Вартанов А.В. Механизмы семантики: человек – нейрон – модель. // Нейрокомпьютеры: разработка, применение. № 12, 2011 с. 54-64.
4. Вартанов А.В., Вартанова И.И. Что такое эмоции? 4-х мерная сферическая модель аспектов переживания, выражения, восприятия и обозначения эмоций. // В сб. Культурно исторический подход и проблема творчества: Материалы вторых чтений памяти Л.С.Выготского, / Под ред. Е.Е.Кравцовой, В.Ф.Спиридонова, Ю.Е.Кравченко.- М. (РГГУ, фонд им. Л.С.Выготского), 2003, с. 13-29.
5. Вартанов А.В., Вартанова И.И. Эмоции, мотивация, потребность в филогенезе психики и мозга. Вест.Моск.Ун-та. Сер. 14 Психология. 2005. N3 с.20-35.
6. Вартанов А.В., Виденеева Н.М. Четырехмерная сферическая модель эмоций и дистанционный речевой контроль состояния человека. Тезисы докладов рабочей группы «Влияние информационных технологий на национальную безопасность» 4-й Ежегодной Конференции Консорциума ПрМ «Построение стратегического сообщества через образование и науку». Москва, 25-27 июня 2001 г., 35 с.

7. Виденеева Н.М., Хлудова О.О., Вартанов А.В. Эмоциональные характеристики звучащего слова. // Журн. высш. нерв. деят., 2000. т.50 вып. 1, с. 29-43.
8. Вилюнас В.К., Основные проблемы психологической теории эмоций, Психология эмоций. М., 1984. с 3-26.
9. Вундт В. Психология душевных волнений. / Психология эмоций. Тексты. М., 1984. с 48-63.
10. Дарвин Ч. Выражение эмоций у человека и животных. Сочинения / Под ред. Н.П.Павловского. М.-Л., 1940. Т. 5.
11. Златоустова Л.В. Типы эмфатического ударения в русском литературном языке. / Общеуниверситетский сборник, т.117, 1957, с. 107-111.
12. Измайлов Ч.А., Коршунова С.Г., Соколов Е.Н. Сферическая модель различения эмоциональных выражений схематического лица человека. // Журн. высш. нерв. деят. т.49, 1999. вып. 2. с. 186-199
13. Калюжный, М.В. Система реабилитации слабовидящих на основе настраиваемой сегментарной модели синтезируемой речи: дис. …канд. тех. наук / М.В. Калюжный. СПб., 2009.
14. Михайлов В.Т., Златоустова Л.В. Измерения параметров речи. // Радио и связь. М.1987.
15. Никишкян Э.А. Типология интонации эмоциональной речи. Киев-Одесса., 1986.
16. Перервенко, Ю.С. Исследование инвариантов нелинейной динамики речи и принципы построения системы аудиоанализа психофизиологического состояния: дис. …канд. тех. наук / Ю.С. Перервенко. Таганрог, 2009.
17. Розалиев, В.Л. Моделирование эмоциональных реакций пользователя при речевом взаимодействии с автоматизированной системой: дис. …канд. тех. наук / В.Л. Розалиев. Волгоград: ВГТУ, 2009.
18. Романенко Р. Ю. Вейвлет-анализ речевых сигналов. // Успехи современной радиоэлектроники. Зарубежная радиоэлектроника. № 12, 2010 с.51-54
19. Сидоров К.В., Филатова Н.Н. Анализ признаков эмоционально окрашенной речи // Вестник Тверского государственного технического университета. – Тверь, 2012. – Вып. 20. – С. 26-31.
20. 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.
21. Симонов П.В. Лекции о работе головного мозга: потребностно-информационная теория высшей нервной деятельности. М.: Наука, 2001. -95 с.
22. Симонов П.В. Эмоциональный мозг. М.: Наука, 1981.
23. Соколов Е.Н. Сферическая модель интеллектуальных операций. // Психол. журн., 2001, том 22, №3, с. 49-56
24. Соколов Е.Н., Вайтнявичюс Г.Г. Нейроинтеллект: от нейрона к нейрокомпьютеру. М.: Наука, 1989. 238 с.
25. Соловьева, Е.С. Методы и алгоритмы обработки, анализа речевого сигнала для решения задач голосовой биометрии: дис. …канд. тех. наук / Е.С. Соловьева. М., 2008.
26. Станиславский К.С. Моя жизнь в искусстве. М., 1959.
27. Старченко И.Б., Перервенко Ю.С., Борисова О.С., Момот Т.В. Методы нелинейной динамики для биомедицинских приложений // Известия ЮФУ. Технические науки. 2010. – № 9 (110). – С. 42-51.
28. Филатова Н.Н., Сидоров К.В. Модель интерпретации знака эмоций по естественной речи // Известия ЮФУ. Технические науки Тематический выпуск. Т. 134, № 9, 2012 с. 39-45.
29. Фролов М.В., Милованова Г.Б. Речевой сигнал как показатель функционального состояния человека-оператора. // Биомедицинская радиоэлектроника. № 6, 2009, с. 49-53.
30. Шпикуляк И. Ему не все равно: смартфоны смогут различать эмоции. // Ноосфера, IT и электроника. http://noos.com.ua/ru/post/3104 / (дата обращения 06 декабря 2012 16:19)
31. 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.

32. Хроматиди, А.Ф. Исследование психофизиологического состояния человека на основе эмоциональных признаков речи: дис. …канд. тех. наук / А.Ф. Хроматиди. Таганрог, 2005.
33. Animal language – Режим доступа: http://animalang.biggo.ru/prakticheskoe_znachenie_/ (дата обращения: 12.02.2013). Практическое значение машины автомата.
34. Chen, Y.T. A study of emotion recognition on mandarin speech and its performance evaluation: Ph. D. dissertation / Y.T. Chen. Tatung, 2008.
35. Intel Developer Centers // http://developer.intel.com; http://www.intel.com/content/www/us/en/search.html?keyword=SPL+ http://www.intel.com/content/www/us/en/search.html?context=767188&tab=767189&keyword=IPPS (дата обращения: 12.12.2013).
36. 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.
37. Osgood C.E., Suci G.J. & Tannenbaum P.H. The measurement of meaning. Urbana. University of Illinois Press. 1957.
38. Schlosberg H.S. A scale for the judgerment of facial expressions // Exsperimental Psychology, 1941, P. 497-510.
39. Siging, W. Recognition of human emotion in speech using modulation spectral features and support vector machines: master of science dissertation / W. Siging. Kingston, 2009.