Автоматическое определение изменений эмоционального состояния по речевому сигналу

Лукьяница А.А.,
кандидат физико-математических наук

Шишкин А.Г.,
кандидат физико-математических наук

Настоящая работа посвящена проблеме создания автоматической компьютерной системы, позволяющей определить изменение эмоционального состояния диктора на основе речевого сигнала. Системы подобного типа относятся к системам распознавания образов, которые обычно состоят из двух основных частей: первая должна выделять из речи наиболее информативные характерные признаки, а вторая (классификатор) — на основе выделенных признаков принимать решение об изменениях в эмоциональном состоянии человека. В данной работе речевой сигнал описывается 211 характерными признаками, что позволило в итоге получить достоверность распознавания состояния диктора с точностью 97.2%. Классификатор был построен на основе метода опорных векторов. Проведённые исследования показали, что число признаков может быть сокращено до 57 самых важных, что приводит к снижению точности лишь на 1.1%.

This paper reports results in development of a computer system that explores the emotional state of a human by his speech. The system consists of two parts: speech features extraction and human state classification. We classify speech into two emotional states based on 211 features with total classification error less than 3%. The classification is performed by support vector machine method. We show that the number of speech features can be reduced to 57 most important ones, thus increasing classification error by 1.1% only.

Ключевые слова: эмоциональное состояние, частота основного тона, форманты, кепстральные коэффициенты, линейное предсказание, метод опорных векторов

1. Введение

При изменении эмоционального состояния в человеческом организме происходят сложные процессы, которые в конечном итоге находят отражение в виде мышечных сокращений, в том числе и в голосовом тракте. Это даёт возможность бесконтактного определения эмоционального состояния человека по изменениям в системе речеобразования. Системы подобного типа, как правило, состоят из двух основных частей, первая из которых должна выделять из речи наиболее информативные характерные признаки, а вторая — на основе выделенных признаков принимать решение, является ли данный образец речи спокойным или соответствует стрессовым изменениям в состоянии человека. Первую часть будем называть системой выделения характерных признаков, а вторую — распознающей, или классифицирующей, системой.

Многочисленные исследования [1—7] показали, что в состоянии даже лёгкого волнения у человека меняется частота основного тона и нескольких первых формант, изменяется спектральный состав речи, повышается энергия высокочастотных компонент, увеличиваются громкость и темп речи, появляется вибрация, растягиваются гласные, а также происходят другие изменения, которые могут быть описаны в математической форме. В качестве характерных признаков наиболее часто используются следующие величины: F_0 — частота основного тона, а также её дисперсия σ_F^2 и вариабельность δ_F (т.е. слабые изменения); F_1, F_2 и F_3 — частота первых трёх формант; I — интенсивность речи и её вариабельность δ_I; J — вибрация (дрожание) голоса; P_I — расположение пиков в звуковой волне; TEO (Teager Energy Operator) [3, 4].

Перечисленные характерные признаки использовались различными исследователями для определения психоэмоционального состояния человека и, судя по описанным результатам, «зарекомендовали себя с положительной стороны». Каждый из этих признаков может изменяться в диапазоне от 1% до 10% при изменении состояния испытуемого, и в совокупности они могут позволить построить систему с высоким уровнем достоверности. Однако это не исключает использования других признаков, которые могут оказаться эффективными именно в российских условиях, с учётом особенностей русской речи.

Настоящая работа организована следующим образом. В следующем разделе описаны выбранные нами характерные признаки и методы их нахождения. Сначала описывается техника отделения речи от пауз, а затем рассматриваются способы вычисления признаков, основанных на определении частоты основного тона, значениях трёх первых формант, а также на вычислении кепстра. Далее приведено краткое описание классифицирующей системы, основанной на методе опорных векторов. В последнем разделе описываются результаты численных экспериментов, направленных на сокращение числа признаков. В заключении сформулированы основные результаты, полученные в настоящей работе.

2. Выделение характерных признаков

Для выделения характерных признаков речевого сигнала на первом этапе его обработки необходимо отделить речь от пауз. При этом осуществляющие указанную операцию методы должны работать в реальном времени. Следовательно, необходимо использовать такие характеристики речевого сигнала, которые являются довольно простыми, но в то же время позволяют надёжно находить начало и конец слова.

2.1. Определение частоты основного тона

Одним из основных характерных признаков, использующихся практически при всех видах анализа речевого сигнала, является частота основного тона и различные производные от неё параметры. Частота основного тона — это та частота, с которой колеблются голосовые связки человека во время произнесения вокализованных звуков. Вследствие того, что данная величина играет чрезвычайно важную роль при получении окончательного результата разрабатываемой системы, необходимо использовать метод, позволяющий определять её с высокой степенью точности. К настоящему моменту существует большое число различных методов вычисления частоты основного тона [8—11]. Наиболее распространёнными являются методы, основанные на использовании автокорреляционной функции и функции нормированной перекрёстной корреляции. Однако простое использование указанных методов приводит во многих случаях к неправильным результатам (например, удвоению истинной частоты основного тона или получению ненулевых значений для невокализованных фрагментов речи). Поэтому нами был использован алгоритм оптимального выбора значения частоты основного тона из имеющихся возможных значений, основанный на использовании метода динамического программирования. Это позволило избежать огромного числа вычислений, неизбежно возникающих при простом переборе всех существующих временных траекторий частоты основного тона, и получать надёжные результаты.

Обозначим речевой сигнал, полученный с частотой дискретизации F_s, как $x(n)$, где $n = 0,1,2,...$ Тогда нормированная функция перекрёстной корреляции определяется следующим выражением:

$$\phi_i(k) = \frac{\sum_{j=m}^{m+n-1} x(j)x(j+k)}{\sqrt{E(m)E(m+k)}},\ k = 0,\ldots,K-1;\ m = iz;\ i = 0,\ldots,M-1, \quad (1)$$

где K — максимальное значение задержки k, i — индекс очередного речевого сегмента (окна), M — число таких сегментов, $z = tF_s$, t — длина сегмента, E — энергия сигнала;

$$E(m) = \sum_{l=m}^{m+n-1} x^2(l). \quad (2)$$

Величина ϕ всегда принадлежит отрезку $[-1;1]$. При этом $\phi_i(k)$ близка к 1 для задержек k, кратных истинному значению периода основного тона, вне зависимости от наличия или отсутствия быстрых изменений сигнала x.

При наличии низкочастотного шума будут получаться большие значения корреляции для всех задержек в искомом диапазоне. Это будет приводить, в том числе, к тому, что невокализованные фрагменты речевой волны будут классифицироваться как вокализованные. Для устранения данной проблемы будем вычитать из значений сигнала его среднее значение в рассматриваемом окне i, т.е.

$$s_i(j) = x(m+j) - \mu_i,\ m = iz;\ j = 0,\ldots,n+K-1, \tag{3}$$

где

$$\mu_i = \frac{1}{n}\sum_{j=m}^{m+n-1} x(j). \tag{4}$$

Так как число точек в окне n и максимальная задержка K пропорциональны частоте дискретизации F_s, число необходимых арифметических действий для вычисления $\phi_i(k)$ пропорционально F_s^2. Для определения частоты основного тона приходится много раз вычислять нормированную функцию перекрёстной корреляции, что даже при использовании быстрого преобразования Фурье приводит к неприемлемо большим затратам вычислительных ресурсов. Поэтому воспользуемся состоящей из двух этапов процедурой, число действий в которой составляет $O(F_s)$. Сначала осуществим прореживание сигнала с частотой F_{ds}, равной

$$F_{ds} = \left]F_s/(4F_{0_{\max}})\right[, \tag{5}$$

где $]\ [$ обозначает округление до ближайшего целого, а $F_{0_{\max}}$ — максимально возможное значение частоты основного тона.

На первом этапе нормированная функция перекрёстной корреляции для прореженного сигнала вычисляется для задержек $F_{ds}/F_{0_{\max}} \le k \le K$. При этом находится максимальное значение $\phi_{\max}$. Все значения $\phi_i(k)$, превышающие $\delta_1 \cdot \phi_{\max}$, запоминаются в качестве возможных кандидатов для определения частоты основного тона. Для более точной оценки положения локальных максимумов корреляционной функции и их амплитуд используется параболическая интерполяция по трём значениям, определяющим пики при частоте дискретизации F_{ds}. Если число таких пиков превышает N_1, то они упорядочиваются в порядке убывания, а потом отбираются первые N_1 пиков.

На втором этапе $\phi_i(k)$ вычисляется с использованием исходного речевого сигнала, но только для семи значений задержки в окрестности каждого из отобранных на первом этапе локальных максимумов корреляционной функции. Аналогичным образом определяется значение $\phi_{\max}$ и все значения $\phi_i(k)$, превышающие $\delta_1 \cdot \phi_{\max}$. Затем они опять упорядочиваются в порядке убывания, отбираются первые N_1 пиков — и наконец осуществляется параболическая интерполяция для уточнения положения и амплитуды локальных максимумов.

После того как в каждом окне найдены возможные значения частоты основного тона, необходимо отобрать из них одно единственно верное. Для этого воспользуемся методом динамического программирования, предложенным Р. Беллманом [12] для эффективного решения оптимизационных задач. Обозначим через I_i число состояний в окне i $(1 \le I_i \le N_1 + 1)$. Для каждого окна это число будет равно количеству возможных значений частоты основного тона (вокализованные состояния) плюс одно невокализованное состояние. Пусть C_{ij} — это значение j-го локального максимума ϕ в окне i, а $L_{i,j}$ — соответствующее ему значение задержки.

Определим стоимость назначения окну i вокализованного состояния с периодом $L_{i,j}/F_s$ следующим образом:

$$d_{i,j} = 1 - C_{i,j}(1 - \beta L_{i,j}),\quad 1 \le j \le I_i, \tag{6}$$

а стоимость назначения окну i невокализованного состояния — как

$$d_{i,j} = d_0 + \max_j C_{i,j}, \tag{7}$$

где

$$\beta = \frac{\beta_0}{\left(F_s / F_{0_{\max}}\right)}. \tag{8}$$

Величина β_0 позволяет уменьшать значимость больших задержек, чтобы предпочтительным оказывался выбор задержек с меньшим значением. Очевидно, что такой выбор функций стоимости даёт преимущество значениям $C_{i,j}$, близким к единице, и небольшим величинам задержки для вокализованных сегментов, а для невокализованных — значениям $C_{i,j}$, близким к нулю. Параметр d_0 контролирует вероятность назначения окну вокализованного состояния.

Стоимость перехода Δ от состояния j в окне i к состоянию k в следующем окне, при условии что оба этих состояния были вокализованными, выражается как

$$\Delta_{i,j,k} = G_1 \cdot \min\left[\xi_{j,k}, \left(G_2 + \left|\xi_{j,k} - \ln 2\right|\right)\right], \tag{9}$$

где

$$\xi_{j,k} = \left|\ln \frac{L_{i,j}}{L_{i-1,k}}\right|, \quad 1 \le j < I_i; \quad 1 \le k < I_{i-1}. \tag{10}$$

В том случае, когда текущее и последующее состояния являются невокализованными, имеем:

$$\Delta_{i,I_i,I_{i-1}} = 0. \tag{11}$$

Если же состояния для текущего и последующего окна отличаются, то в случае перехода от вокализованного к невокализованному получим

$$\Delta_{i,I_i,k} = H_1 + H_2 \cdot S_i + H_3 \cdot r_i, \quad 1 \le k < I_{i-1}, \tag{12}$$

а для перехода от невокализованного к вокализованному —

$$\Delta_{i,j,I_{i-1}} = H_1 + H_2 \cdot S_i + H_3 / r_i, \quad 1 \le j < I_i. \tag{13}$$

Здесь

$$r_i = \frac{R(i,h)}{R(i-1,h)}, \tag{14}$$

$$R(i,h) = \sqrt{\frac{\sum_{j=0}^{J-1} \left(W_j s_{j+m+h}\right)^2}{J}}, \quad m = iz, \tag{15}$$

где W — окно Хэминга длиной $J = 00.3F_s$:

$$W_j = 0.54 - 0.46\cos\left(\frac{2\pi j}{J}\right), \tag{16}$$

h — величина, контролирующая расстояние между центрами текущего и предыдущего окна для вычисления $R(i,h)$; S является функцией, обратной расстоянию Итакуры D_{IT} [13]:

$$S_i = \frac{0.2}{D_{IT}(i,i-1) - 0.8}. \tag{17}$$

При этом порядок линейного предсказания выбирается как

$$p = 2 + \rceil F_s/1000 \lceil, \tag{18}$$

где $\rceil\ \lceil$ обозначает округление до ближайшего целого.

Обратимся теперь к анализу выражений для функций стоимости (12)—(13). Такой выбор стоимости переходов между состояниями обеспечивает её убывание в том случае, когда спектр сигнала подвержен быстрым изменениям, как в случае между границами вокализованных классов. Положительная константа H_1 является штрафом за изменение вокализованного состояния, вне зависимости от изменений в речевом сигнале. Это является отражением того факта, что смена вокализованных состояний в речевом сигнале происходит довольно редко.

Оптимальная траектория находится следующим образом:

$$D_{i,j} = d_{i,j} + \min_{k \in I_{i-1}} \left[D_{i-1,k} + \Delta_{i,j,k} \right], \quad 1 \le j \le I_i \tag{19}$$

с начальными условиями

$$D_{0,j} = 0, \quad 1 \le j \le I_0; \quad I_0 = 2. \tag{20}$$

Для каждого состояния в каждом окне необходимо хранить индексы для обратного прохода:

$$q_{i,j} = k_{\min}, \tag{21}$$

где $k_{\min}$ — индексы k, минимизирующие $D_{i,j}$ в каждом окне.

Значение частоты основного тона в окне i определяется как

$$F_{0_i} \frac{F_s}{L_{i,j}}, \tag{22}$$

где значения j — это те величины, для которых достигается глобальный минимум D.

В проводимых нами расчётах использовались следующие значения параметров, фигурирующих в формулах (5)—(13):

$F_s = 22050$ Гц, $F_{ds} = 2205$ Гц, $F_{0_{\max}} = 450$ Гц, $\delta_1 = 0.3$, $N_1 = 20$, $d_0 = 0$,

$\beta_0 = 0.3$, $G_1 = 0.2$, $G_2 = 0.35$, $H_1 = 0.005$, $H_2 = 0.5$, $H_3 = 0.5$.

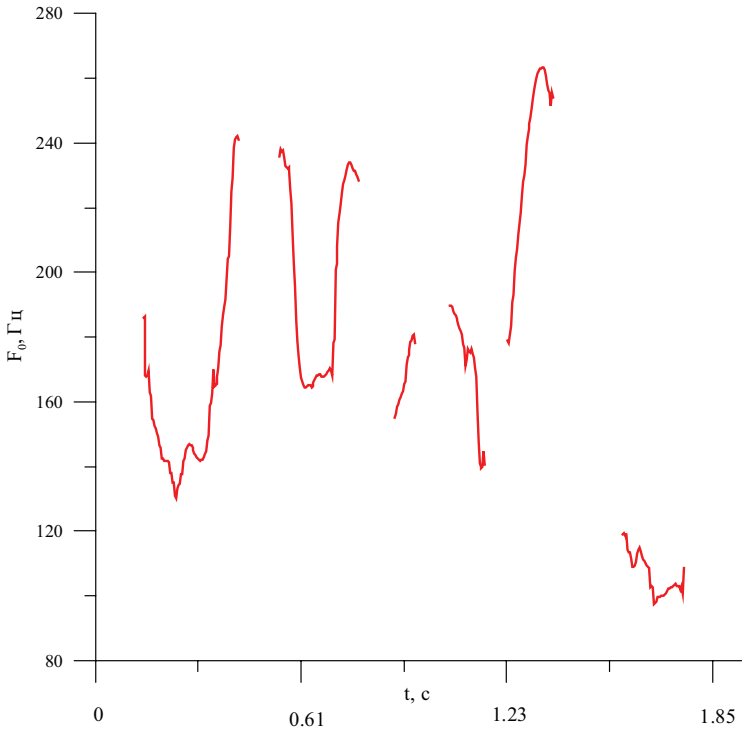

Рис. 1. Траектория частоты основного тона для отдельного фрагмента речевого сигнала

На рис. 1 представлены результаты применения описанного выше алгоритма для определения частоты основного тона для отдельного фрагмента речевого сигнала.

2.2. Определение значений первых трёх формант

Для определения необходимых нам первых трёх формант мы воспользуемся методом линейного предсказания [14]. В большинстве случаев, после того как коэффициенты линейного предсказания посчитаны, буфер коэффициентов дополняется единицей слева и нулями до степени числа 2 справа и подаётся на вход быстрому преобразованию Фурье (БПФ). Дополнение нулями нужно для повышения точности, так как при длине буфера меньше определённого значения (а в данном случае в буфере первоначально всего 20 коэффициентов) БПФ даёт очень неточные результаты. В результате применения быстрого преобразования Фурье и последующего преобразования составляющих спектра получается огибающая спектра сигнала, максимумы которой представляют собой формантные частоты.

Однако данный метод является весьма неточным и не позволяет надёжным образом определять значения формантных частот. Поэтому мы применяли другой способ нахождения формант, который основан на нахождении комплексных корней полинома [15]:

$$A(z)=1-\sum_{k=1}^{p}\alpha_k z^{-k}=\prod_{k=1}^{p}\left(1-z_k z^{-1}\right), \quad (23)$$

где z — вектор коэффициентов линейного предсказания $\left[-1 \quad a_1 \quad a_2 \ldots a_p\right]$.

Данные корни могут быть найдены с помощью метода Лагерра, имеющего кубическую сходимость для изолированных корней и линейную сходимость для кратных корней. Текущая итерация берётся в качестве корня, когда значение полинома при этой итерации меньше, чем вычисленная граница ошибок округления при нахождении значения полинома для этой итерации. Степень исходного полинома уменьшается, когда найден вещественный корень или пара комплексных корней, после чего итерационный процесс применяется к полиному уменьшенной степени [16].

Найденные корни многочлена могут быть представлены в следующем виде:

$$z_k=\exp\left(\left[-\pi b_k+j2\pi F_k\right]/F_s\right), \quad (24)$$

где через b_k, F_k и F_s обозначены частотный диапазон, центральная частота k-й форманты и частота дискретизации соответственно. Так как α_k являются действительными, все комплексные корни имеют сопряжённые им, т.е. если (b_k, F_k) — корень, то $(b_k, -F_k)$ также представляет собой корень. Все b_k всегда положительны, так как для устойчивого предсказателя корни должны лежать внутри единичного круга $(|z_k|<1)$. Действительные корни не учитываются при нахождении формант, а комплексные корни упорядочиваются по возрастанию положительных F_k. Кроме того, исключаются корни, у которых частотная полоса превышает 200 Гц.

Указанный метод позволяет достаточно надёжно определять значения формант, однако в ряде случаев, как, например, при переходе от вокализованных к невокализованным участкам, а также при смене звучания фонем, возможно некорректное нахождение формант. Поэтому был использован метод динамического программирования, позволяющий учесть временные профили формант и тем самым значительно повысить точность их определения.

Введём базовые значения первых трёх формант F_{ni}, i = 1,2,3: F_{n1} = 500 Гц, F_{n2} = 1500 Гц, F_{n3} = 2500 Гц. Наша задача заключается в выборе среди N формант на протяжении K окон. В каждом окне k существует N_k способов отнести возможных кандидатов к определённым формантам:

$$N_k=\frac{n!}{(n-N)!N!}, \quad (25)$$

где n — число кандидатов формант в предыдущем окне, а N — рассматриваемое число формант.

Форманты выбираются из числа кандидатов по принципу минимальной стоимости, которая зависит от локальной стоимости, стоимости изменения частоты и стоимости перехода. Локальная стоимость λ_{kl} l-го назначения в k-ом окне зависит от частотного диапазона b_{kln} и отклонения от базового значения формант F_{nn}:

$$\lambda_{kl} = \sum_{n=1}^{N} \left[\beta_n b_{kln} + \nu_n \mu_n \frac{|F_{kln} - F_{nn}|}{F_{nn}} \right], \quad (26)$$

где β_n — стоимость увеличения частотного диапазона для n-й форманты, ν_n — вероятность того, что данное окно является вокализованным, а μ_n определяет стоимость отклонения от базового значения n-й форманты.

Стоимость изменения частоты ξ_{kljn} между l-ым назначением в k-м окне и j-ым назначением в $(k-1)$-м окне для n-й форманты выражается следующим образом:

$$\xi_{kljn} = \left[\frac{F_{kln} - F_{k-1jn}}{F_{kln} + F_{k-1jn}} \right]^2. \quad (27)$$

Стоимость перехода δ_{klj} определяется как

$$\delta_{klj} = \psi_k \sum_{n=1}^{N} \alpha_n \xi_{kljn}, \quad (28)$$

где α_n задаёт относительную стоимость изменения частоты между окнами для n-й форманты. Член ψ_k контролирует степень непрерывности траектории форманты:

$$\psi_k = \frac{\Re_k}{\max_{i \in K} \Re_i}, \quad (29)$$

где $\Re_k$ — среднеквадратичное значение сигнала в k-м окне.

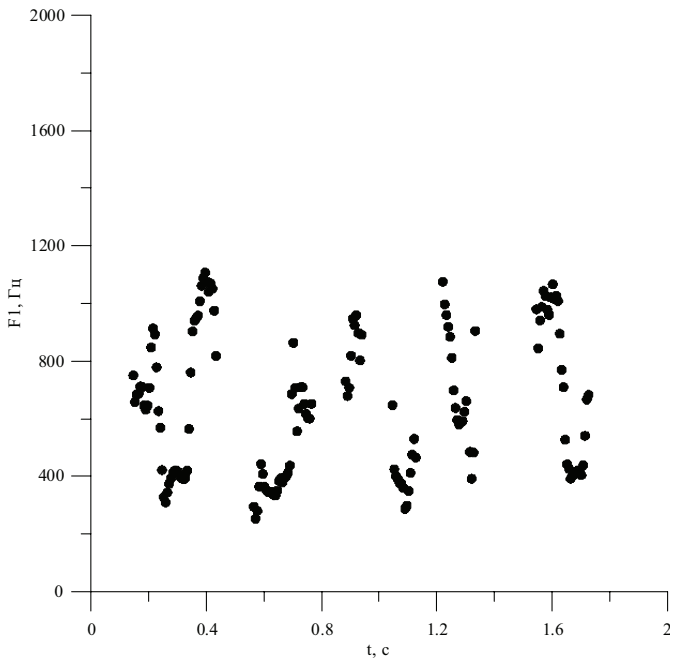

Рис. 2. Изменение первой форманты во времени для того же фрагмента речи, что и на Рис. 2.

Окончательно минимальная общая стоимость выбора кандидатов частоты форманты на протяжении K окон с N_k возможными соответствиями (назначениями) в каждом окне выражается следующим образом:

$$C = \sum_{k=1}^{K} \min_{l \in N_k} D_{kl} . \quad (30)$$

Здесь стоимость D_{kl} задаётся как

$$D_{kl} = \lambda_{kl} + \min_{j \in N_{k-1}} \kappa_{klj}, \quad (31)$$

где κ_{klj} — стоимость перехода от j-го назначения в $(k-1)$-м окне к l-му назначению в k-м окне:

$$\kappa_{klj} = \delta_{klj} + D_{k-1\,j} . \quad (32)$$

На рис. 2 показаны результаты применения указанного алгоритма для вычисления первой форманты для отдельного фрагмента речевой волны. Из рисунка следует, что значения первой форманты лежат в пределах от 300 до 1200 Гц.

2.3. Признаки, основанные на вычислении кепстра

Существует немало частотных характеристик речевого сигнала, однако устойчивых к внешним шумовым помехам и позволяющих адекватно описывать данный отрезок сигнала среди них — единицы. Одними из наиболее успешно применяемых на практике являются кепстральные коэффициенты нелинейного масштаба [17]. Они определяются как действительный кепстр кратковременного сигнала, полученный из Фурье-преобразования данного сигнала. Отличие от действительного кепстра состоит в том, что используется нелинейная частотная шкала. В процессе построения программной системы нами были использованы первые 14 коэффициентов.

Амплитуда речевого сигнала существенно изменяется во времени. В частности, амплитуда невокализованных сегментов речевого сигнала значительно меньше амплитуды вокализованных сегментов. Подобные изменения амплитуды хорошо описываются с помощью функции кратковременной энергии сигнала. Кроме того, при изменении функционального состояния диктора происходит перераспределение энергии сигнала из одних частотных диапазонов в другие. Это приводит к тому, что среди признаков речевого сигнала, использующихся для выявления различных эмоциональных состояний, должны быть энергетические характеристики. Для получения энергии в определённых частотных диапазонах необходимо сначала вычислять свёртку исходного сигнала с полосовыми фильтрами, а затем уже подсчитывать энергию.

В качестве фильтра мы использовали так называемый Windows-Sinc-Blackman фильтр, имеющий следующую импульсную характеристику:

$$h(i) = K \frac{\sin(2\pi f_c (i - M/2))}{i - M/2} \left[0.42 - 0.5 \cos\left(\frac{2\pi i}{M}\right) + 0.08 \cos\left(\frac{4\pi i}{M}\right) \right], \quad i = 0, \ldots, M$$

где K — константа нормировки, которая выбирается так, чтобы $\sum_{i=0}^{M} h(i) = 1$. Отметим, что M должно быть чётным числом.

3. Используемые характерные признаки

Для разделения всех речевых сигналов на два класса, соответствующих нормальному и стрессовому состояниям, с помощью методов, описанных в предыдущих разделах, были выделены 211 характерных признаков. В результате для каждого отдельного речевого сигнала получается вектор, размерность которого равна 211.

Для проведения испытаний качества работы разработанных методов и алгоритмов, реализованных в программной системе, необходимо наличие базы данных (БД), в которой хранятся речевые сигналы испытуемых, находящихся в различных эмоциональных состояниях. Проведённый анализ показал, что к настоящему моменту существует лишь одна такая русскоязычная база данных — RUSLANA (RUSsian LANguage Affective speech), созданная в университете Мейкай в Японии [18]. Она состоит из 3660 предложений, произнесённых 49 женщинами и 12 мужчинами в возрасте от 16 до 28 лет, для которых русский язык является родным. К сожалению, указанная БД недоступна для использования.

Среди англоязычных БД, безусловно, выделяется SUSAS (Speech Under Simulated and Actual Stress), собранная в течение нескольких лет в Колорадском университете в г. Боулдер (США) [19]. БД содержит записи 32 человек в возрасте от 22 до 76 лет, в том числе записи переговоров пилотов военных вертолётов, выполняющих боевые задания и, естественно, находящихся в стрессовых условиях. Отметим также БД из Северной Ирландии Belfast Natural Database [20] и из Массачусетского технологического института (США) [21].

Существует довольно большое число эмоциональных БД на немецком, японском и голландском языках [22—25]. В большинстве случаев в указанных БД рассматриваются следующие виды эмоций (приведены в порядке убывания частоты их использования): злость, печаль, радость, страх, отвращение, удивление, скука, стресс, презрение и др. Для свободного использования фактически доступны только две БД. Одна из них, имеющая очень небольшой размер, создана в Институте электроники и телекоммуникаций в г. Марибор (Словения) и содержит записи на четырёх языках — английском, французском, испанском и словенском [26]. Вторая БД — Berlin Database of Emotional Speech (EmoDB) — является наиболее подходящей для целей нашего исследования [27]. Она состоит из речевых фрагментов, произнесённых 5 мужчинами и 5 женщинами на немецком языке. Каждый из испытуемых произносил 10 фраз (например, «Скатерть лежит на холодильнике» или «Они только что отнесли это наверх и сейчас идут обратно»), пытаясь сымитировать 6 различных эмоциональных состояний: злость, скука, отвращение, беспокойство/страх, радость, печаль. Кроме этого, дикторы произносили фразы нормальным спокойным голосом. Для некоторых эмоций в БД существует несколько вариантов их озвучивания одним и тем же лицом.

Тестирование разработанного программного комплекса осуществлялось на данной БД. Был создан специальный модуль, позволяющий считывать каждый звуковой файл из БД по отдельности и записывать выделенные характерные признаки в единый файл. После этого все вектора характерных признаков были разделены в соотношении 70/30. Первая часть была использована для обучения распознающей системы, работающей по методу опорных векторов. Оставшиеся вектора служили в качестве тестовой последовательности. Использовался метод 10-кратной перекрёстной валидации, при котором меняются последовательности, служащие в качестве обучающих и тестовых. Рассматривалось два класса: нормальное состояние и отклонение от нормы, куда попали все 6 имеющихся эмоциональных состояний. Особую трудность настройке распознающей системы придаёт тот факт, что в рассматриваемой БД эмоционально окрашенные файлы составляют порядка 90%. Таким образом, обучающая последовательность содержит значительно больше звуковых файлов, произнесённых в состоянии, отличном от нормального.

4. Классифицирующая система

Разрабатываемая система должна обладать возможностью относить речевой сигнал к одному из двух классов — в зависимости от того, был он произнесён в нормальном состоянии или отличном от него. Для построения подобного классификатора наилучшим образом подходит метод опорных векторов [28—29], позволяющий вычислить оптимальную разделяющую поверхность. Пусть нам известны N примеров $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, где $\boldsymbol{x}_i \in R^n$ — векторы характерных признаков, а $y_i \in \{-1, +1\}$ — переменная, характеризующая эмоциональное состояние диктора: значение $y_i = -1$ соответствует нормальному состоянию, в то время как $y_i = +1$ означает состояние, отличное от нормального. В качестве разделяющей поверхности мы использовали гиперплоскость

$$(\boldsymbol{w} \cdot \boldsymbol{x}) - b = 0, \quad \boldsymbol{w} \in R^n, \; b \in R, \tag{34}$$

поскольку она обладает наилучшими экстраполирующими свойствами по сравнению с нелинейными поверхностями. Соответствующий классификатор имеет следующий вид:

$$f(\boldsymbol{x}) = \operatorname{sgn}((\boldsymbol{w} \cdot \boldsymbol{x}) - b), \tag{35}$$

т.е. в соответствии со знаком правой части вектор характерных признаков $\boldsymbol{x}$ относится к тому или иному множеству.

Значения коэффициентов $(\boldsymbol{w}, b)$ гиперплоскости находятся на основе обучающего множества в результате максимизации следующего функционала Лагранжа [28]:

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{36}$$

при дополнительных ограничениях

$$\alpha_i \geq 0, \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \quad i = 1, \ldots, N. \tag{37}$$

Здесь α_i — множители Лагранжа. Величина коэффициента b определяется из условий Куна-Таккера:

$$\alpha_i\left[y_i\left((\boldsymbol{w}\cdot\boldsymbol{x}_i)-b\right)-1\right]=0,\quad i=1,\ldots,N. \tag{38}$$

Процесс нахождения коэффициентов $(\boldsymbol{w},b)$ обычно называется настройкой или обучением системы.

В результате настройки точность распознавания состояния говорящего, т.е. отнесения к одному из двух возможных классов, составила 97.2%. Необходимо отметить, что это является одним из лучших на сегодняшний день результатов по эмоциональной классификации речевых сигналов. Типичным для большинства имеющихся двухклассовых систем распознавания является уровень порядка 70—80%.

5. Сокращение числа параметров

Для описанной системы существуют приложения, для которых очень важными характеристиками является время обработки речевого сигнала, а также объём используемой памяти: например, если система будет встроена в мобильный телефон, то это позволит во время разговора определять эмоциональное состояние собеседника. Для повышения производительности можно попытаться уменьшить число обрабатываемых параметров (разумеется, если это приведёт к незначительному снижению качества распознавания). Кажется естественным, что в первую очередь нужно отбросить характеристики сигнала, которые оказывают наименьшее влияние на результат. Как показали проведённые исследования [30], наилучшим образом значимость параметров характеризуется величиной его вариации, которая определяется следующим образом.

Приведём все компоненты векторов признаков к диапазону [0,1] и выберем k-ю компоненту: $k=1,\ldots,n$. Для каждого из заданных векторов признаков $\boldsymbol{x}_i$, $i=1,\ldots,N$, вычислим расстояние d_i^k до найденной разделяющей гиперплоскости вдоль k-й компоненты:

$$d_i^k=\left|x_H^k-x_i^k\right|, \tag{40}$$

где

$$x_H^k=\frac{b-\sum_{j=k}^{n}w_j x_i^j}{w_k}, \tag{41}$$

а x_i^k — k-я компонента i-го вектора признаков $\boldsymbol{x}_i$. Вариация ν_k k-й компоненты определяется по формуле

$$\nu_k=\frac{1}{N}\sum_{i=1}^{N}d_i^k. \tag{42}$$

Чем меньше величина вариации, тем больше значимость соответствующего признака, поэтому степень значимости p_k можно выразить следующим образом (для удобства — в процентах):

$$p_k = \frac{\nu_k}{\sum_{k=1}^{n} \nu_k} \times 100\% . \qquad (43)$$

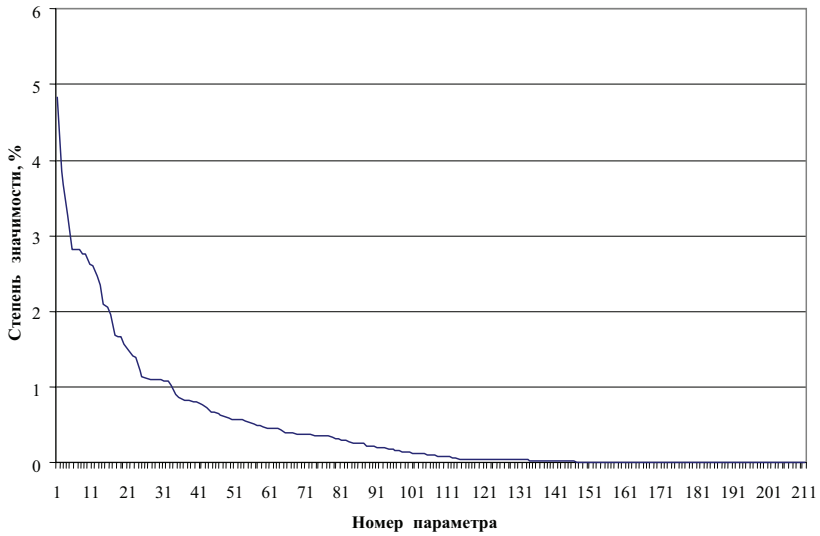

Рис. 3. Распределение степени значимости параметров

На рис. 3 приведено распределение степени значимости используемых 211 параметров речевого сигнала. Видно, что влияние на результат более чем половины параметров не превосходит 0.5%. Это обстоятельство позволяет сократить число параметров. Нами были оставлены параметры, значимость которых превосходит следующие пороговые значения: 0.5%, 1%, 1.5%, 2%. В следующей таблице (см. табл. 1) приведена зависимость точности определения эмоционального состояния дикторов от числа используемых параметров.

Таблица 1

Зависимость точности распознавания от числа параметров

Важность параметров , %	Число параметров, шт.	Точность распознавания, %
>0	211	97,2
>0.5	57	96,1
>1.0	34	95,9
>1.5	29	89,6
>2.0	15	86,7

Из таблицы видно, что сокращение числа параметров с 211 до 57 приводит к снижению точности лишь на 1.1%, а всего при 15 параметрах точность остаётся достаточно приемлемой для многих приложений. В таблице 2 перечислены наиболее важные для рассматриваемой задачи 15 параметров.

Таблица 2

Степень значимости наиболее важных параметров

Наименование параметра	Степень значимости, %
Медиана 12-го кепстрального коэффициента	4,84
Интерквартильный размах 7-го кепстрального коэффициента	3,86
Интерквартильный размах 6-го кепстрального коэффициента	3,68
Минимальное значение 10-го кепстрального коэффициента	3,27
Интерквартильный размах 5-го кепстрального коэффициента	2,82
Среднее значение 12-го кепстрального коэффициента	2,81
Медиана 7-го кепстрального коэффициента	2,81
Медиана 3-го кепстрального коэффициента	2,77
Интерквартильный размах 13-го кепстрального коэффициента	2,76
Минимальное значение 6-го кепстрального коэффициента	2,62
Минимальное значение 13-го кепстрального коэффициента	2,60
Среднее значение 1-го кепстрального коэффициента	2,47
Интерквартильный размах 1-го кепстрального коэффициента	2,34
Интерквартильный размах 4-го кепстрального коэффициента	2,09
Интерквартильный размах 7-го кепстрального коэффициента	2,05

Интересно, что все указанные параметры вычисляются на основе кепстра, что является весьма удачным обстоятельством с точки зрения повышения быстродействия системы. К сожалению, дальнейшее сокращение числа параметров приводит к резкому снижению достоверности классификации, что позволяет утверждать, что для рассматриваемого множества данных данное число параметров является минимально допустимым.

6. Заключение

Исследования, проведённые авторами при выполнении настоящей работы, показали высокую эффективность метода определения изменений в эмоциональном состоянии человека на основе анализа речевого сигнала. Достигнутая точность в 97.2% позволяет использовать такую систему для вынесения экспертных заключений, например, в бесконтактных «детекторах лжи», которые могут использоваться в финансовых учреждениях при выдаче кредитов. Сильно усечённый вариант системы, использующий на порядок меньшее число признаков, характеризующих речевой сигнал, требует незначительных вычислительных ресурсов, обеспечивая при этом точность 86.7%, чего вполне достаточно для интегрирования в бытовую технику, например, в мобильные телефоны.

Литература

1. *L.Rothkrantz et al*. Voice Stress Analysis. Text, Speech and Dialogues, ISBN 3-540-23049-1, *Lecture Notes in Artificial Intelligence*, P. 449—456, Springer, Berlin-Heidelberg-New York, 2004.
2. *O-W.Kwon et al*. Emotion Recognition by Speech Signals. In: *Proc. Intern. Conf. EUROSPEECH 2003*, Geneva. P. 125—128, 2003.
3. *G.Zhou, J.H.L. Hansen, J.F. Kaiser*. «Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator», In: *Proc. IEEE Inter. Conf. on Acoustics, Speech, Signal Processing*, vol.I. P. 549—552, Seattle, 1998.
4. *Zhou G.; Hansen J.H.L.; Kaiser J.F*. «Methods for Stress Classification: Nonlinear TEO and Linear Speech Based Features». In: *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.4. P. 2087—2090, 1999.
5. *B.D. Womack, J.H.L. Hansen*. «Classification of Speech Under Stress using Target Driven Features», *Speech Communication, Special Issue on Speech Under Stress*, vol.20(1-2). P. 131—150, 1996.
6. *G. Zhou, J.H.L. Hansen and J.F. Kaiser*. «Nonlinear Feature Based Classification of Speech under Stress», *IEEE Transactions on Speech & Audio Processing*, 1997.
7. *M.Sigmund*. «Spectral Analysis of Speech under Stress». *Int. Journal of Computer Science and Network Security*, vol.7. P. 170—172, 2007.
8. W.J.Hess, *Pitch Determination of Speech Signals-Algorithms and Devices*, Springer-Verlag, Berlin, 1983.
9. *L.R.Rabiner, M.J.Cheng, A.E.Rosenberg and A.McGonegal*. «A comparative study of several pitch determination algorithms», *IEEE Trans. On Acoustics, Speech and Signal Processing*, Vol.ASSP-24: 399—413.
10. *Y.Tadokoro, W.Matsumoto and M.Yamaguchi*. «Pitch Detection of Musical Sounds Using Adaptive Comb Filters Controlled by Time Delay», In: *Proc. of the International Conf. on Multimedia and Expo*. P. 109—12, 2002.
11. *A.Cheveigne and H.Kawahara*. «YIN, a Fundamental Frequency Estimator for Speech and Music», *The Journal of the Acoustical Society of America*, vol. 111, Issue 4, pp.1917—30, 2002.
12. *Беллман Р., Энджел Э.* Динамическое программирование и уравнения в частных производных, М.: Мир, 1974, с. 208.
13. *F.Zheng, Z.Song, L.Li, W.Yu, F.Zheng, W.Wu*. The Distance Measure For Line Spectrum Pairs Applied to Speech Recognition, In: *Proc. 5th International Conference on Spoken Language Processing*, Sydney, №. 0171, 1998.
14. *Маркел Дж.Д., Грей А.Х.* Линейное предсказание речи. М.: Радио и связь, 1980. 248 с.
15. *L.Rabiner, B.-H.Juang*. Fundamentals of Speech Recognition. Prentice Hall, 1995. 507 p.
16. *Уилкинсон Дж.Х.* Алгебраическая проблема собственных значений, «Наука», М., 1970.
17. *X.Huang, A.Acero, H.W.Hon*. Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001. 1008 p.
18. *V.Makarova and V.A.Petrushin*. «RUSLANA: A Database of Russian Emotional Utterances», In: *Proc. 2002 Int. Conf. Spoken Language Processing (ICSLP 2002)*, Colorado. P. 2041—2044, 2002.
19. Linguistic Data Consortium (LDC), http://www.ldc.upenn.edu/.
20. *E.Douglas-Cowie, R.Cowie and M.Schroeder*. «A New Emotion Database: Considerations, Sources and Scope», In: *Proc. ISCA (ITWR) Workshop Speech and Emotion: A conceptual framework for research*, Belfast. P. 39—44, 2000.
21. *R.Fernandez and R.W.Picard*. «Modeling Drivers' Speech Under Stress», In: *Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research*, Belfast, 2002.
22. *F.Schiel, S.Steininger, U.Turk*. «The Smartkom Multimodal Corpus at BAS», In: *Proc. Lang. Resources and Evaluation*, Canary Islands, 2002.
23. *B.Wendt and H.Scheich*. «The Magdeburger Prosodie-Korpus», In: *Proc. Speech Prosody Conf. 2002*, Aix-en-Provence. P. 699—701, 2002.
24. *Y.Niimi, M.L. Kasamatu, T.Nishimoto and M.Araki*. «Synthesis of Emotional Speech Using Prosodically Balanced VCV Segments», In: *Proc. 4th ISCA tutorial and Workshop on research synthesis*, Scotland, 2001.
25. *S.J.L. Mozziconacci and D.J. Hermes*. «A study of intonation patterns in speech expressing emotion or attitude: production and perception», *IPO Annual Progress Report 32*, P .154—160, IPO, Eindhoven, 1997.
26. Emotional Speech, http://www.elektronika.uni-mb.si/eSpeech/speech.html.

27. *F.Burkhardt, A.Paeschke, M.Rolfes, W.Sendlmeier and B.Weiss*. A Database of German Emotional Speech, In: *Proc. Intern. Conf. Interspeech*, Lissabon, 2005.
28. *V.N.Vapnik*. The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science), 2nd edition. New York, Springer-Verlag, 1999. 304 p.
29. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1973. 416 с.
30. *A.A.Lukianitsa, F.M.Zhdanov and F.S.Zaitsev*. «Analysis of ITER Operation Mode Using the Support Vector Machine Technique for Plasma Discharge Classification», Plasma Physics and Control Fusion, v.50, №6. P. 14, 2008.

Шишкин Алексей Геннадиевич

Старший научный сотрудник факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова, кандидат физико-математических наук.

Область научных интересов: математическое моделирование, распознавание образов, адаптивные методы обработки сигналов различного вида, обработка изображений, распознавание речи, применение современных информационных технологий в научных исследованиях.

Автор свыше 90 научных работ, в том числе одного учебника и одной монографии.

Лукьяница Андрей Александрович

Старший научный сотрудник факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, кандидат физико-математических наук.

Область научных интересов: вычислительные методы, распознавание образов, нелинейная оптимизация, искусственные нейросети, генетические алгоритмы, скрытые модели Маркова, обработка изображений, распознавание речи.

Автор более 80 научных работ, включая один учебник и три монографии.