УДК 004.52

ПРОГРАММНАЯ ОБОЛОЧКА РАСПОЗНАВАНИЯ КОМАНД В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

А. С. Четкин, С. А. Запрягаев

Воронежский государственный университет

Поступила в редакцию 30.05.2016 г.

Аннотация. В работе представлена программная оболочка, описание алгоритмов и процедур для распознавания ограниченной группы слов на основе скрытых моделей Маркова (СММ). Распознавание осуществляется в режиме реального времени с использованием функции нормального распределения вероятностей наблюдаемых событий. Приведены результаты исследования точности распознавания в зависимости от числа скрытых состояний, числа фонем в слове и числа кепстральных коэффициентов в векторах признаках.
Ключевые слова: распознавание речи, скрытые модели Маркова.
Annotation. The paper presents the program, description of algorithms and procedures for recognition of a limited group of words based on the hidden Markov model (HMM). The recognition is the process in the real time using a normal probability distribution function of observed events. The study results of recognition accuracy are presented as a functions of the numbers of hidden states, the number of phonemes in the word and the number of cepstral coefficients in the feature vectors.
Keywords: speech recognition, hidden Markov model.

ВВЕДЕНИЕ

Естественный ввод информации в цифровые системы - актуальная задача для различных технических приложений. Примером таких приложений является система распознавания речи, которая может быть использована для голосового управления, голосовой идентификации, голосового перевода, компрессии речи и т. п.

В настоящее время предложено большое число методов и подходов для построения систем распознавания речи в зависимости от их назначения: диктор-независимые системы, on-line и off-line системы, системы, распознающие слитную речь или отдельные команды, и т. п. В данной работе на основе построения вектор-признаков и скрытых моделей Маркова представлена on-line программная оболочка распознавания и выполнен анализ влияния выбора моделей Маркова на точность распознавания. В целом разработанную систему распознавания речи можно представить в виде последовательности процедур, представленных на рис. 1.

1. ЗАПИСЬ И ОЦИФРОВКА СИГНАЛА

Работа блоков записи и оцифровки сигнала (рис. 1) основывается на том, что речи человека соответствует, в основном, частотный диапазон колебаний аналогового сигнала $s(t)$ от 200 до 4000 Гц. Известно, что оцифрованный сигнал $s_i \quad i \in 0,1,2,\ldots,N-1$ по теореме Котельникова может быть восстановлен однозначно и без потерь при частоте дискретизации, вдвое большей максимальной частоты в сигнале. Поэтому для дискретного описания человеческой речи достаточно ее представление в диапазоне до 8 кГц. Так как при оцифровке звука важным параметром является амплитудное квантование, то установлено, что при использовании квантования с разрядностью в 16 бит погрешности квантования остаются для слушателя практически незаметными. В результате данная разрядность амплитуды колебаний рассматривается в работе в качестве приемлемой величины.

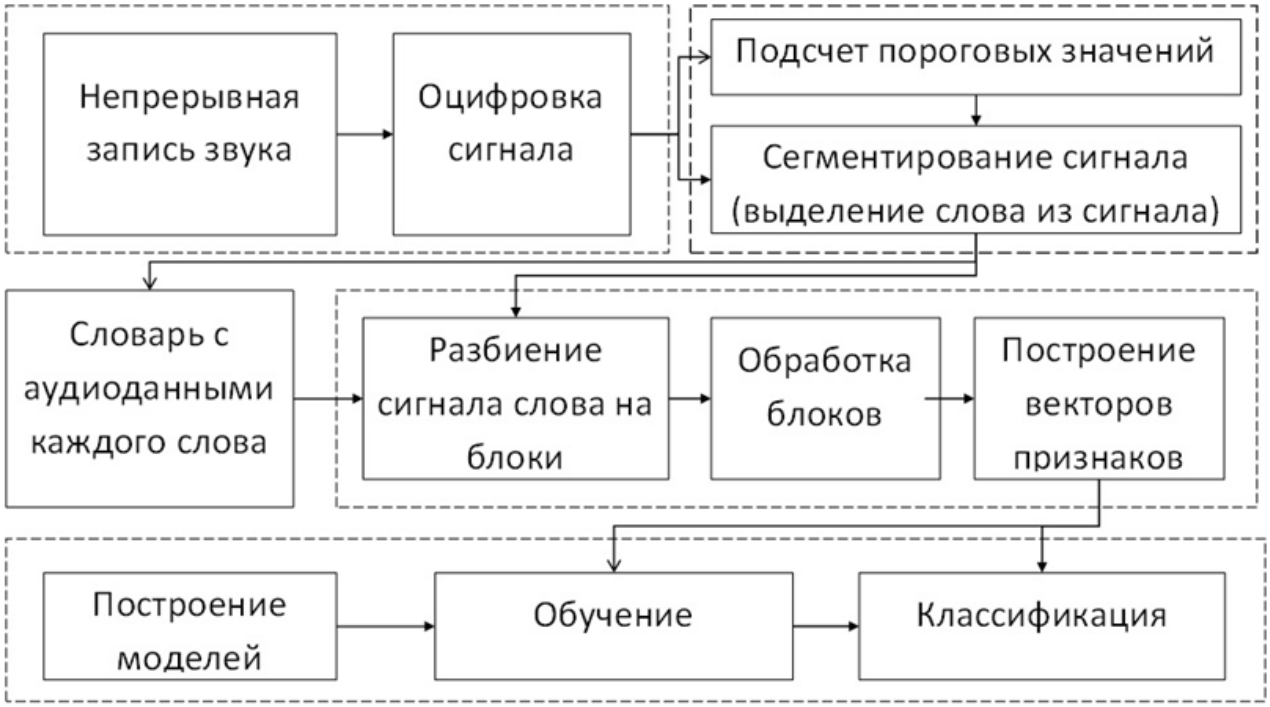

Рис. 1. Блок-схема системы распознавания речи

2. ВЫДЕЛЕНИЕ СЛОВА ИЗ СИГНАЛА

Оцифрованный сигнал требуют наличия системы выделения слова из непрерывного потока звуковой информации (рис. 1). При этом необходимо учесть возможное изменение условий, в которых производится запись речи. Наличие шума и вариация интенсивности звуковых сигналов являются источниками возможных ошибок. В связи с этим система определения начала и конца слова должна обеспечить их минимальное влияние на результат.

В данной работе решение задачи определения начала и конца слова в сигнале (VAD – voice activity detection) в режиме реального времени основано на анализе величин четырех различных характеристик сигнала: среднее значение «огибающей» сигнала, нормированное число переходов оцифрованного сигнала через нуль, отношение интенсивностей сигнал-шум и максимальная длительность «тишины».

Первая характеристика основана на определении среднего значения модуля оцифрованной «огибающей» $H_n(s)$ сигнала s_i, $i = 0,1,\ldots N-1$:

$$\frac{2}{N}\sum_{n=0}^{[N/2]}\left|H_n(s)\right|, \qquad (1)$$

где $H_n(s)$ определено комбинацией преобразований Фурье вида:

$$H(s) = F^{-1}(F(s)\cdot 2\theta) \qquad (2)$$

Здесь F и F^{-1} – прямое и обратное дискретное преобразование Фурье, соответственно, $\theta(k)=1$ для $k = 0,1,\ldots,[N/2]$ и $\theta(k)=0$ для $k > [N/2]$, $[x]$ – целая часть x. Очевидно, что у речевого сигнала значение параметра $E(\mathrm{s})$ больше среднего значения параметра $E(\mathrm{s})$ у «тишины» сигнала или постороннего слабого шума, что и может быть использовано, как одно из условий, при выделении речевой фазы сигнала.

Второй характеристикой сигнала, используемой для выделения речевой фазы, является нормированное число переходов оцифрованного сигнала через нуль:

$$Z(s) = \frac{1}{N-1}\sum_{k=1}^{N-1}\frac{1}{2}\left|\,\mathrm{sgn}(\mathrm{s}_k) - \mathrm{sgn}(s_{k-1})\,\right|. \qquad (3)$$

Здесь $\mathrm{sgn}(x)$ – знак числа x. В общем случае значение параметра $Z(s)$ лежит в диапазоне от 0 до 1 и позволяет оценить частотный диапазон сигнала. Чем ближе $Z(s)$ к единице,

тем выше частота сигнала $s(t)$. Очевидно, что для речи $Z(s)$ будет меньше, чем, например, для высокочастотного шума, т. к. максимум энергии в спектре речи находится в диапазоне частот от 200 до 1500 Гц. В среднем значения $Z(s)$ для речи человека находятся в диапазоне от 0.05 до 0.15. Для шума с частотой выше 2000 Гц данное значение будет выше 0.5, поэтому при определении данного параметра для высокочастотного шума (без речевого сигнала) необходимо его значение инвертировать для сохранения возможности распознавания речевой фазы. В этом случае Z изменяется на значение $1-Z$.

Третьей характеристикой, используемой в настоящей работе для выделения слова, является отношение интенсивностей спектральных плотностей сигнал-шум. Для определения спектральной плотности сигнала s используется дискретное Фурье-преобразование F с окном Хеминга w

$$X_k(s) = F(s_i \cdot w(i)), \qquad (4)$$

где $w(i) = 0.5\left(1-\cos\left(\frac{2\pi i}{N-1}\right)\right)$. Если в качестве сигнала анализируется оцифрованный шум d_i без речевого сигнала, то спектральная плотность шума определяется аналогичным выражением

$$D_k(s) = F_1(d_i \cdot w(i)). \qquad (5)$$

В результате параметр $S(s)$, характеризующий отношение сигнал-шум, имеет вид

$$S(s) = 10\log_{10}\left(\frac{1}{N}\sum_{k=0}^{N-1}\left|\frac{X_k(s)}{D_k(s)}\right|^2\right). \qquad (6)$$

В разработанной системе для определения параметров $E(s)$, $Z(s)$, $S(s)$ звуковой сигнал разбивается на блоки по 20 мс во временном интервале со сдвигом окна, равным 20 мс. В каждом из блоков вычисляются параметры $E_i(s)$, $Z_i(s)$, $S_i(s)$, где i – номер блока. В процессе распознавания данные параметры сравниваются со значениями предварительно установленных порогов для этих величин: $\overline{E}(s)$, $\overline{Z}(s)$, $\overline{S}(s)$. В настоящей работе пороговые значения вычисляются по следующему алгоритму:

1. Записывается звук (шум) длительностью 2 секунды, в котором отсутствует речевой сигнал. Записанный сигнал разбивается на 100 блоков длительностью по 20 мс каждый.

2. В каждом блоке записанного шума вычисляются $E_i(s)$, $Z_i(s)$, $S_i(s)$, $i \in 1,2,\ldots,k$. Здесь $k = 100$ – число блоков

3. Из набора значений в 100 блоках для каждого из трех параметров $E_i(s)$, $Z_i(s)$, $S_i(s)$ определяются их максимальные значения.

$$\overline{E} = \max\left(E_i(s)\right);\ \overline{Z} = \max\left(Z_i(s)\right);$$

$$\overline{S} = \max\left(S_i(s)\right).$$

Найденные максимальные значения $\overline{E}$, $\overline{Z}$, $\overline{S}$ рассматриваются в качестве пороговых для параметров $E_i(s)$, $Z_i(s)$, $S_i(s)$ при определении фазы речи в каждом блоке реального сигнала.

Обычно в литературе [1] предлагается считать анализируемый сигнал речевым, если значение более чем одной из трех характеристик превышает пороговое значение в VAD-функции

$$VAD_i = \theta\left(\delta_{1i} + \delta_{2i} + \delta_{3i} - 1\right), \qquad (7)$$

где $\delta_{1i} = \theta\left(E_i(s) - \overline{E}\right)$, $\delta_{2i} = \theta\left(Z_i(s) - \overline{Z}\right)$, $\delta_{3i} = \theta\left(S_i(s) - \overline{S}\right)$, а $\theta(x)$ – функция Хевисайда.

В процессе распознавания блоки звукового сигнала длиной 20 мс последовательно подаются на вход функции VAD и если значение данной функции равно «1» на протяжении как минимум 10 блоков, т. е. в течение 200 мс, то данный участок сигнала считается словом. Однако в некоторых словах могут возникать паузы длиной больше, чем 10 блоков. Например, в слове «четыре» между буквами «ч» и «т» присутствует такая пауза, поэтому алгоритм на основе VAD разобьет данное слово на две части, как показано на рис. 2а. Таким образом, если не учитывать возможные паузы в словах, то существенно уменьшится точность определения начала и конца слова. Поэтому в данной работе введен дополнительный параметр – *максимальная длительность тишины*, который учитывает возможность появления относительно длинной паузы между звуками в словах. На основании анализа группы слов экспериментально было выбрано максимальное значение длины тишины, равное 100 мс, или 5 блокам. Если в течение пяти блоков значение функции VAD вновь

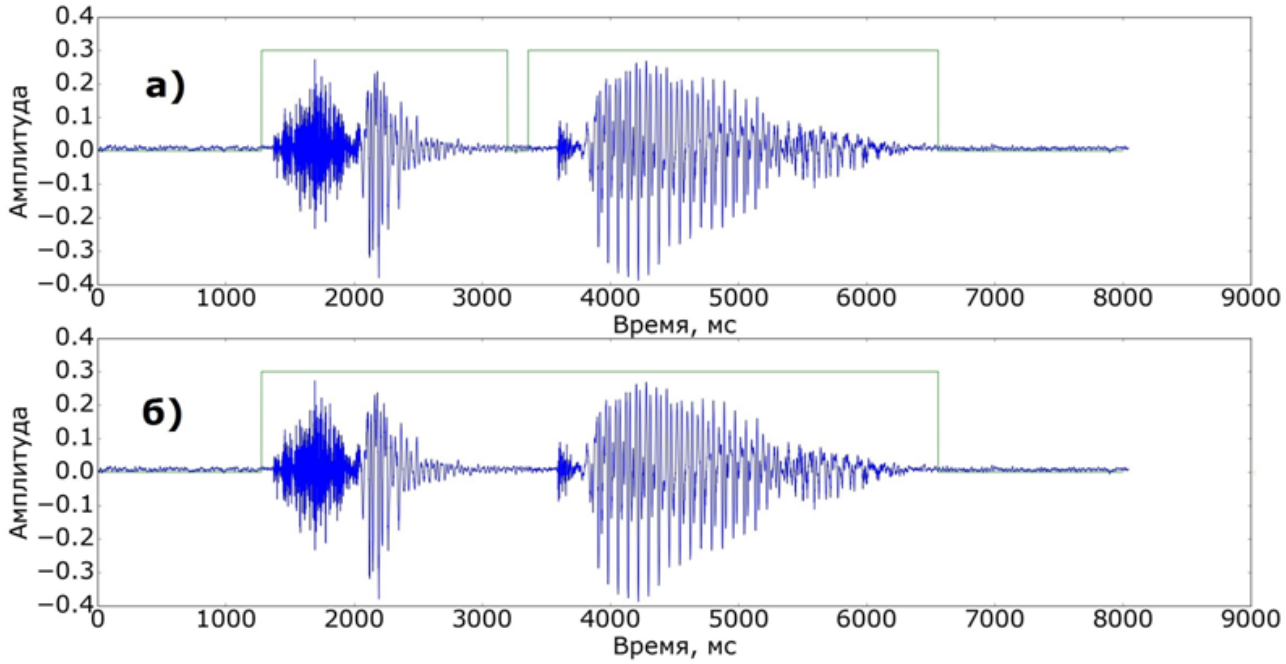

Рис. 2. а) Результаты определение начала и конца слова «четыре» без учета параметра «максимальная длительность тишины», б) с учетом данного параметра

изменилось на «1», то предыдущие участки считаются также принадлежащими речевой фазе и определяют начало слова. Соответственно условием, указывающим на достижение конца слова, считается количество блоков тишины, превышающих значение 5. На рис. 2б изображен результат применения разработанного алгоритма определения начала и конца слова с учетом максимальной длительности тишины.

3. ПОСТРОЕНИЕ ВЕКТОРОВ ПРИЗНАКОВ

Выделение речевых участков сигнала, в которых присутствуют слова, позволяет провести предварительную подготовку словаря распознаваемых слов на основе процедуры извлечения векторов признаков у слов, входящих в состав словаря. Так как диктор может произносить слова с различной интенсивностью и на разном расстоянии от микрофона, то используемые при регистрации значения амплитуд сигнала будут различаться даже для одних и тех же слов. Поэтому, прежде чем приступить к извлечению векторов признаков, необходимо нормировать полученные амплитуды к выбранному интервалу. В настоящей работе нормировка амплитуды сигнала S_i осуществляется в диапазоне от –1 до 1 по формуле:

$$a_i = 2 \cdot \frac{s_i - \min(s)}{\max(s) - \min(s)} - 1, \quad i = 1, 2, \ldots, N. \quad (8)$$

Здесь S_i – оцифрованное значение амплитуды сигнала, N – число дискретных значений сигнала, $\max(s)$ и $\min(s)$ – максимальное и минимальное алгебраические значения сигнала в выделенной речевой фазе соответственно.

Вычисленный нормированный сигнал разбивается на последовательность блоков, состоящих из K точек оцифрованного сигнала, сдвинутых по времени на фиксированную величину в P точек. При частоте дискретизации 8000 Гц размер блока во временном интервале составляет 32 мс, сдвиг – 10 мс. Выбранный сдвиг окна блока обеспечивает величину пересечения блоков при одном сдвиге на 68.75 %. Число блоков Q, на которое разбивается сигнал, при заданной длине сигнала N определяется по формуле:

$$Q = \left[\frac{N - K}{P} \right] + 1, \quad (9)$$

где $[x]$ – целая часть числа x.

После разбиения сигнала на блоки полученные данные можно представить в следующем виде: $a_k^{(q)}$, где k – номер точки сигнала в

блоке $k = 0,\ldots,K-1$, а q – номер блока $q = 0,\ldots,Q-1$. Векторы-признаки вычисляются в каждом блоке q.

Для получения векторов-признаков используется подход, близкий к физиологическому восприятию звуков ухом человека. Особенностью восприятия звука человеком является тот факт, что в полосе частот звуковых колебаний до 1000 Гц субъективное восприятие удвоения частоты почти линейно совпадает с физическим увеличением частоты в два раза. Но в диапазоне частот выше 1000 Гц зависимость субъективного восприятия удвоения частоты становится близкой к логарифмической [2]. Явление нелинейного психофизического восприятия звука позволяет дифференцированным образом строить векторы признаков в различных частотных диапазонах, в то время как стандартная оцифровка сигнала осуществляется равномерно. С целью учета отмеченной особенности восприятия звука часто используется переход от частотного диапазона представления сигнала к так называемой мел-шкале. При этом значение частоты f, заданное в герцах, преобразуется в безразмерное значение «высоты» сигнала, обозначаемой m или мел, по формуле [3]:

$$\mathrm{mel}(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right). \qquad (10)$$

Соответственно обратное преобразование от мел к частотной шкале имеет вид:

$$F(\mathrm{mel}) = 700(e^{(\mathrm{mel}/2595)} - 1). \qquad (11)$$

Для имитации особенности восприятия звука человеческим ухом в настоящей работе при построении векторов признаков используются так называемые мел-кепстральные коэффициенты [4].

Для нахождения мел-кепстральных коэффициентов (внутри каждого блока q) на первом этапе выполняется оконное преобразование Фурье оцифрованного речевого сигнала $a_k^{(q)}$ с окном Хэмминга $w(k)$ по формуле:

$$A_n^{(q)} = \sum_{k=0}^{[K/2]-1} w(k)\, a_k^{(q)} \exp\left(-i \cdot \frac{2\pi}{K-1} k \cdot n\right), \quad n = 0,1,\ldots,\left[\frac{K}{2}\right]-1. \qquad (12)$$

Окно Хемминга применяется для того, чтобы уменьшить растекание спектра, т. к. исходный сигнал разбивается на блоки малой длины, что влечет за собой уменьшение разрешения частотной сетки.

Учитывая частоту дискретизации, использованную в настоящей работе, равную 8000 Гц, при переходе к мел-шкале получим максимальное значение, равное $\mathrm{mel}(4000) = 2146$. В диапазоне значений $\mathrm{mel} = 0,\ldots,2146$ строится набор из M штук равных треугольных фильтров, равномерно расположенных на мел-оси (рис 3). Точки основания треугольника z_m $(m = 0,1,\ldots M+1)$ на частотной шкале определены по формуле:

$$z_m = F\left(m \cdot \frac{\mathrm{mel}(4000)}{M}\right) \text{Гц.}, \quad m = 0,1,\ldots M+1. \qquad (13)$$

Для задания правила построения M штук треугольных фильтров введем вспомогательный номер Δ_m, $m = 0,1,\ldots M+1$ в соответствии с определением $\Delta_m = z_m \cdot K / 8000$. Здесь K – число точек в блоке, 8000 – выбранная частота дискретизации. В результате «m»-ый треугольный фильтр $(m = 1,2,\ldots,M)$ в каждой точке дискретизации $k = 0,1,\ldots,K$ определен следующим образом:

$$H_m[k] = \begin{cases} 0, & \text{для } k < \Delta_{m-1} \\ \dfrac{k - \Delta_m}{\Delta_m - \Delta_{m-1}}, & \text{для } \Delta_{m-1} \le k \le \Delta_m \\ \dfrac{\Delta_{m+1} - k}{\Delta_{m+1} - \Delta_m}, & \text{для } \Delta_m \le k \le \Delta_{m+1} \\ 0, & \text{для } k > \Delta_{m+1} \end{cases} \qquad (14)$$

Данные фильтры позволяют вычислить лог-энергию спектра $E^{(q)}[m]$ каждого фильтра в блоке с номером q по формуле:

$$E^{(q)}[m] = \sum_{k=0}^{K-1} \ln\left[\left|A_n^{(q)^2}\right| H_m[k]\right], \quad m = 1,2,\ldots,M, \; q = 1,2,\ldots,Q. \qquad (15)$$

Соответственно M штук мел-кепстральных коэффициентов $c^{(q)}[m]$, где $m = 1,2,\ldots,M$ в каждом блоке q вычисляются с использованием дискретного косинус-преобразования для $E^{(q)}[m]$:

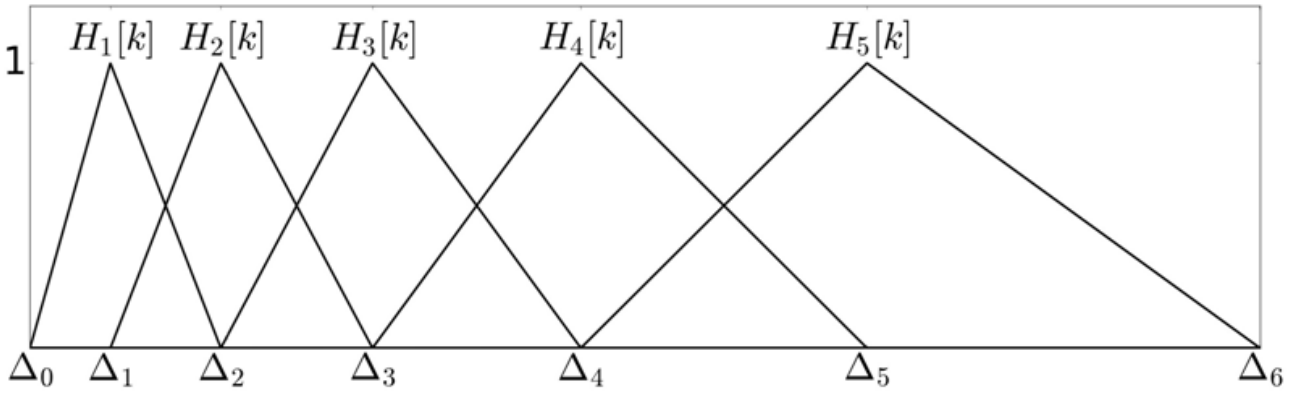

Рис. 3. Положение треугольных фильтров на частотной оси

$$c_k^{(q)} = \sum_{m=1}^{M} E^{(q)}[m] \cos\left(\pi \frac{k(m-1/2)}{M+1}\right), \quad (16)$$

$$k = 1, 2, \ldots, M.$$

Для дальнейшего изложения обозначим M -мерный вектор кепстральных коэффициентов речевого блока с номером q через $\vec{C}^{(q)} \in \left(c_1^{(q)}, c_2^{(q)}, \ldots, c_{M=26}^{(q)}\right)$.

4. КЛАССИФИКАЦИЯ ВЕКТОРОВ ПРИЗНАКОВ С ПОМОЩЬЮ СКРЫТЫХ МОДЕЛЕЙ МАРКОВА

В качестве классификатора полученных векторов-признаков в настоящей работе применяются скрытые модели Маркова, в которых наблюдаемые события считаются результатом перехода от одного скрытого состояния к другому. Состояние описывается только наблюдаемыми событиями. В случае распознавания речи наблюдаемые события – это последовательность векторов кепстральных коэффициентов в блоках, на которые разбит сигнал. Однако число скрытых состояний и функции распределения вероятностей наблюдения коэффициентов неизвестны, и нельзя сказать, в каком из состояний находилась модель при наблюдении текущего события.

В Марковской модели с дискретными отсчетами времени система определяется в любой момент времени одним из N различных состояний. В дискретные моменты времени система переходит из одного состояния в другое (или остается в прежнем состоянии) с определенной вероятностью, образуя матрицу $A \in a_{ij}$ вероятностей переходов из состояния i в состояние j. Задание матрицы A и вероятностей нахождения системы в начальный момент времени π образуют Марковскую модель $\lambda = (A, \pi)$.

Описание *скрытой* модели Маркова (СММ) дополняется вероятностями наблюдаемых событий, которые порождаются состояниями системы на основе заданной функции вероятности наблюдения события. В такой модели о системе можно судить только по набору наблюдаемых событий. Вероятности наблюдения события i в состоянии j образуют матрицу $B \in b_{ij}$. В результате скрытые модели Маркова с дискретными отсчетами времени описываются тремя параметрами $\lambda = (A, B, \pi)$. Если речь рассматривать как последовательность перехода из одного состояния в другое (например, последовательность перехода по фонемам слова), то скрытые Марковские модели можно рассматривать в качестве инструмента распознавания.

Для решения задачи распознавания речи большое распространение получили СММ с непрерывными распределениями вероятностей b_{ij}. Главное их преимущество перед СММ с дискретными значениями вероятностей в том, что с входными данными не надо производить векторное квантование, которое может вносить искажения. В скрытых моделях Маркова с дискретными распределениями вероятностей каждое событие должно быть заранее предопределено. Если в процессе работы с данной моделью наблюдаемые события не встречаются в такой модели, то необходимо для данных событий найти ближайшие из представленных (т. е. квантовать наблюдаемые события) в модели. Кван-

тование может быть осуществлено, например, с помощью k-ближайших соседей. Поэтому чем меньше в модели предопределённых событий, тем меньше точность, а в наблюдаемые события будут вноситься искажения. Так как наблюдаемые события (векторы признаки – мел-кепстральные коэффициенты) не могут быть описаны конечным множеством векторов-признаков, то при векторном квантовании часть информации потеряется, что будет приводить к искажению и потерям. А в скрытых моделях Маркова с непрерывными распределениями вероятностей b_{qj} может быть задана как функция плотности распределения или их смесь.

В настоящей работе используются скрытые модели Маркова с нормальным распределением вероятностей наблюдаемых событий. В настоящей работе вероятность наблюдения кепстральных коэффициентов блока q для n-го экземпляра речевого сигнала в состоянии j задается гауссовой функцией распределения вида

$$b_{qj} \equiv \varphi_j\left(\vec{C}^{(q)}(n)\right) = \frac{1}{\left(2\pi\right)^{K/2}\sqrt{\left|\Sigma_j\right|}} \times \\ \times \exp\left[-\frac{1}{2}\left(\vec{Y}_j^{(q)}(n)\right)^T \cdot \Sigma_j^{-1} \cdot \vec{Y}_j^{(q)}(n)\right]. \quad (17)$$

Здесь $\vec{Y}_j^{(q)}(n) \equiv \vec{C}^{(q)}(n) - \vec{\mu}_j$, K – число кепстральных коэффициентов в блоке, Σ_j – матрица ковариации в скрытом состоянии j, $\vec{\mu}_j \in \left(\mu_1^{(j)}, \mu_2^{(j)}, \ldots, \mu_{26}^{(j)}\right)$ – K-мерный вектор средних значений кепстральных коэффициентов, определенный для заданного скрытого состояния в соответствии с приведенными ниже выражениями, n – номер экземпляра слова для обучения. Как следует из выражения (17), для построения функции $\varphi_j\left(\vec{C}^{(q)}(n)\right)$ необходимо выбрать число скрытых состояний модели. Если, например, выбранное число скрытых состояний равно W, то для определения $\vec{\mu}_j$ выбирается объединение Π_{qi} всех кепстральных коэффициентов с заданным номером i $(i = 1,2,\ldots,K)$ имеющихся блоков речевых сигналов Q_1, Q_2,…, Q_p с учетом выбранных «p» штук экземпляров речевых сигналов одного слова. Число блоков в различных экземплярах сигнала может быть различным (в зависимости от условий произношения). Таким образом:

$$\Pi_{qi} = \begin{pmatrix} c_1^{(1)}(1) & \cdots & c_K^{(1)}(1) \\ \vdots & \ddots & \vdots \\ c_1^{(Q_n)}(n) & \ddots & c_K^{(Q_n)}(n) \\ \vdots & \ddots & \vdots \\ c_1^{(Q_p)}(1) & \cdots & c_K^{(Q_p)}(p) \end{pmatrix}. \quad (18)$$

На основе значений матричных элементов Π_{qi} с помощью алгоритма k-средних находятся W кластеров. W-штук центроидов вычисленных кластеров назначаются i-ми значениями вектора средних значений кепстральных коэффициентов $\mu_i^{(j)}$, принадлежащих j-ым состояниям $(j = 1,2,\ldots W)$. Повторяя данную процедуру для всех значений номеров кепстральных коэффициентов $i = 1,2,\ldots,K$ будут найдены W-штук K-мерных векторов средних значений кепстральных коэффициентов, определенных для заданного j-го скрытого состояния. Другими словами, определены W-штук векторов $\vec{\mu}_j \in \left(\mu_1^{(j)}, \mu_2^{(j)}, \ldots, \mu_{26}^{(j)}\right)$, здесь $j = 1,2,\ldots,W$.

Для вычисления параметра Σ в (17) строится матрица ковариации по определению

$$\bar{X}_i = \frac{1}{Q_\Sigma}\sum_{q=1}^{Q_\Sigma}\Pi_{qi},\ i = 1,2,\ldots,26; \\ Q_\Sigma = Q_1 + \ldots + Q_p, \quad (19)$$

где $V_{qi} = \Pi_{qi} - \bar{X}_i$ и $\Sigma = \frac{1}{Q_\Sigma - 1}V^T \cdot V$. Полученная матрица Σ считается матрицей ковариации для всех состояний, т. е. $\Sigma_j \equiv \Sigma$. Размерность каждой матрицы ковариации равна $K * K$, где K – число элементов вектора признака. Число матриц ковариации и средних значений равно числу состояний в модели.

Перед началом обучения модели параметры A и π задаются случайным образом. Полученные вероятности и параметры модели $\lambda = \left(A, \vec{\mu}_j, \Sigma_j, \pi\right)$ используются в EM-алгоритме [5] для пересчета данных параметров.

После обучения всех моделей с помощью EM-алгоритма, выбирается новая последовательность векторов признаков, не участвовавшая в обучении, которая и подается на проверку каждой из обученных моделей. Тогда данная последовательность относится

к тому слову, в модели которой была получена максимальная вероятность. Вероятность принадлежности вычисляется с помощью алгоритма прямого-обратного хода [5].

5. ЭКСПЕРИМЕНТАЛЬНОЕ ПРИМЕНЕНИЕ СММ В ЗАДАЧЕ РАСПОЗНАВАНИЯ РЕЧИ

В рамках данной работы рассмотрена задача распознавания речи на конечном наборе слов. В словаре были выбраны следующие 50 слов и команд: «Вверх», «Вниз», «Влево», «Вправо», «Старт», «Стоп», «Вперед», «Назад», «Разворот», «Открыть», «Закрыть», «Включить», «Выключить», «Удалить», «Сохранить», «Отменить», «Подтвердить», «Набрать», «Выделить», «Вставить», «Печатать», «Компьютер», «Пуск», «Документ», «Файл», «Номер», «Программа», «Изображение», «Аудио», «Видео», «Число», «Место», «Время», «Режим», «Слово», «Буква», «Набор», «Город», «Длина», «Папка», «Ноль», «Один», «Два», «Три», «Четыре», «Пять», «Шесть», «Семь», «Восемь», «Девять».

Для исследования качества распознавания были рассмотрены модели с различным числом скрытых состояний модели. В первом случае число состояний выбиралось на основе числа фонем в выбранном слове. Каждая фонема в слове рассматривалась как отдельное состояние. Например, результат разбиения группы команд из 8 слов на фонемы представлен в табл. 1.

В иных случаях были использованы модели с фиксированным числом скрытых состояний для всех слов словаря без связи с числом фонем в слове (команде). Скрытые «состояния» в модели с тремя состояниями интерпретируются как начало, середина и конец слова. В моделях с большим числом фиксированных состояний слово интерпретируется как последовательная эволюция речевого сигнала от начала до возникновения внутренних характерных структур сигнала и до его затухания. Для всех видов моделей была использована матрица переходов A со следующими ограничениями: $a_{ij} = 0$, для $j < i$ и $j > i + 1$. Данные условия означают, что в модели невозможны переходы через состояния

Таблица 1

Анализ на скрытые состояния по количеству фонем для 8 слов

Слово	Вверх	Вниз	Влево	Вправо	Вперед	Назад	Старт	Стоп
Фонемный состав	[в] [в’] [э] [р] [х]	[в] [н’] [и] [с]	[в] [л’] [э] [в] [а]	[ф] [п] [р] [а] [в] [а]	ф] [п’] [и] [р’] [о] [т]	[н] [а] [з] [а] [т]	[с] [т] [а] [р] [т]	[с] [т] [а] [р] [т]
Число состояний	5	4	5	6	6	5	5	4

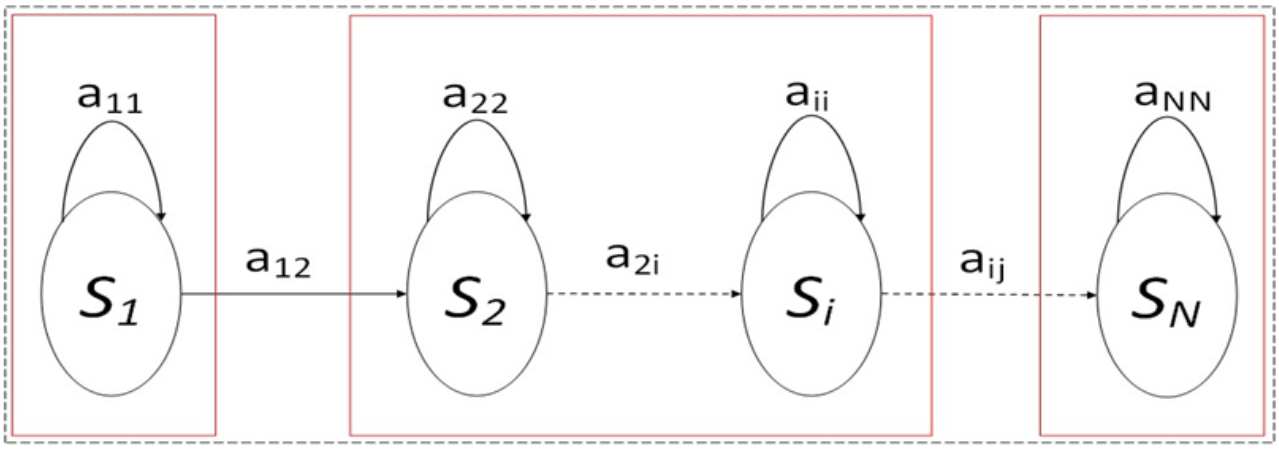

Рис. 4. Общая структура графа состояний в рассмотренных моделях

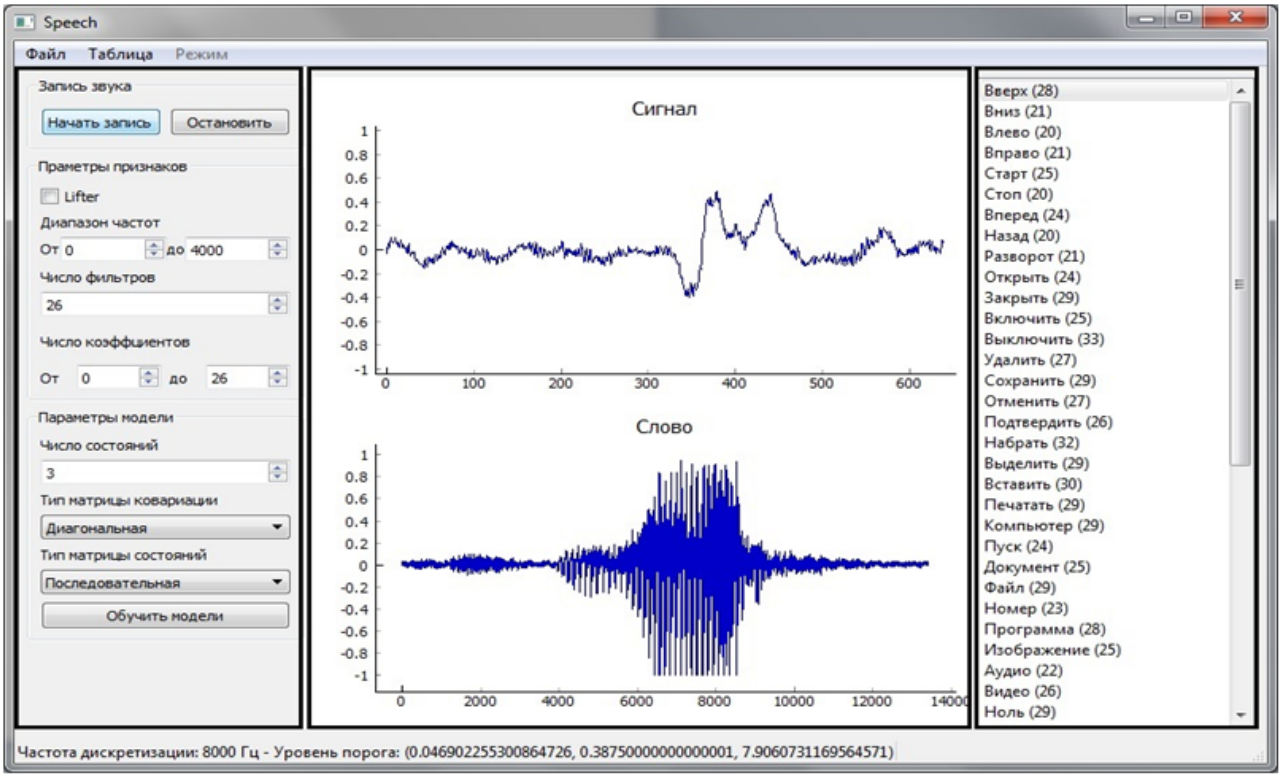

Рис. 5. Окно интерфейса программной оболочки

или возврат к предыдущему состоянию. В виде графа данные ограничения выглядят, как показано на рис. 4.

6. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ СИСТЕМЫ РАСПОЗНАВАНИЯ РЕЧИ

В данной статье разработана программная оболочка для распознавания изолированных команд в режиме реального времени. Программная оболочка была написана на языке Python 3.4 с использованием математических пакетов NumPy и SciPy. В программной оболочке было проведено исследование по распознаванию слов с использованием скрытых моделей Маркова с различным числом состояний. В качестве обучающей выборки для данных моделей выступали 10 слов, произнесенных одним автором, из записанных 50-ти экземпляров слова. Для проведения эксперимента распознавания программная оболочка использовала оставшиеся 40 экземпляров слова. В эксперименте распознавания менялось число заданных скрытых состояний, число кепстральных коэффициентов в блоке, диапазон выборки частот для распознавания и число слов словаря.

Перед началом работы данная программа записывает 2 секунды звуковых данных, которые используются для вычисления порогов для системы определения начала и конца слов. В программной оболочке предусмотрен перерасчет значений порогов, т. к. представленная система детектирования речи не является адаптивной. При нажатии кнопки «Начать запись» запускается система распознавания речи. Когда диктор произносит слово, система детектирует его начало и конец и передает данные для построения векторов признаков. Полученные векторы признаки подаются каждой из моделей, которые вычисляют вероятность совпадения с эталонами. Результат в пользу того или иного слова выбирается исходя из максимального показателя рассчитанной вероятности среди всех моделей. Распознанное слово записывается в таблицу. Остановка работы системы распознавания осуществляется нажатием на кнопку «Остановить». В программе предусмотрена возможность распознавания слова из файла PCM формата «*.wav» путем выбора в меню «Файл» пункта «Распознать из файла».

На рис. 5 изображено главное окно программной оболочки.

Таблица 2

Качество распознавания в зависимости от длины словаря и числа коэффициентов

Число скрытых состояний W		**3**			**5**			**10**		
Число кепстральных коэффициентов K в одном блоке		**13**	**20**	**26**	**13**	**20**	**26**	**13**	**20**	**26**
Число слов в словаре	20	86 %	84 %	82 %	98 %	98 %	96 %	100 %	100 %	94 %
	30	86 %	84 %	82 %	98 %	96 %	94 %	100 %	100 %	94 %
	40	86 %	84 %	82 %	96 %	96 %	94 %	99 %	99 %	92 %
	50	84 %	84 %	82 %	96 %	96 %	94 %	99 %	99 %	92 %

Таблица 3

Точность распознавания группы команд разработанной оболочкой

Команда	Вверх	Вниз	Влево	Вправо	Вперед	Назад	Старт	Стоп
W = число фонем	97 %	94 %	95 %	88 %	88 %	95 %	96 %	98 %
$W = 3$	96 %	97 %	96 %	94 %	94 %	96 %	97 %	98 %
$W = 4$	96 %	90 %	95 %	90 %	88 %	94 %	96 %	95 %
$W = 5$	92 %	90 %	90 %	85 %	84 %	83 %	92 %	94 %

Левая часть окна интерфейса программы на рис. 5 определяет блок задания параметров обучения и записи аудиоданных. Центральные окна отображают графики входных и выходных аудиоданных. Правая часть окна интерфейса программы содержит информацию о словаре и распознанных словах.

ЗАКЛЮЧЕНИЕ

Суммарный результат статистического анализа использования разработанной программной оболочки представлен в табл. 2. В данной таблице приведен процент правильного распознавания слова в зависимости от числа выбранных скрытых состояний W и числа K, выбранных кепстральных коэффициентов в блоке. Статистика набиралась по 40 экземплярам каждого слова, не входившим в обучающую выборку. Результаты в табл. 2 представлены для полосы пропускания в диапазоне 100–3900 Гц и для различных объемов словаря.

Как видно из табл. 2, разработанная система осуществляет распознавание с высокой точностью для словаря до 50-ти слов. При этом наилучший результат соответствует наибольшему числу скрытых состояний с числом кепстральных коэффициентов в диапазоне 15–20. Увеличение объема словаря ведет к последовательному снижению точности распознавания, и при числе слов более 100 точность распознавания падает до 60–70 % . Выбор полосы пропускания незначительно влияет на суммарные статистические данные, приведенные в табл. 2.

В результате проведенного анализа было выявлено, что модель, построенная на числе скрытых состояний W, совпадающих с числом фонем в слове, не является наилучшей. Ряд примеров для выбранной группы команд представлен в табл. 3.

Как видно из табл. 3, для модели с 6 состояниями (слова «Вправо» и «Вперед» имеют 6 фонем) точность распознавания ниже, чем на моделях с меньшим числом заданных скрытых состояний при использовании разработанной модели распознавания.

Время распознавания отдельно произнесённого слова с учетом времени определения начала и конца слова занимает в среднем 100 мс на системе со следующими характеристиками: Intel Core i-3 3.07 ГГц, 4 Гб ОЗУ, Windows 7.

Определенной проблемой представленной программной оболочки является отсутствие очевидного механизма распознавания слов, не представленных в словаре. Для решения данной проблемы используется грубое сравнение вероятности с установленным порогом. Если значение порога больше вычисленной вероятности нового слова, то считается, что данное слово не принадлежит словарю.

В целом представленная программная оболочка решает задачу распознавания отдельных слов в режиме реального времени с ограниченным словарем.

СПИСОК ЛИТЕРАТУРЫ

1. *Ramirez J.* Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding, Michael Grimm and Kristian Kroschel (Ed.) / J. Ramirez, J. M. Gorriz, J. C. Segura. – InTech, 2007 – P. 460.

2. *Норман Д.* Переработка информации у человека / Д. Норман, П. Линдсей; пер. с анг. Д. Лурия. – М. : Мир, 1974. – 550 с.

3. *Homayoon B.* Fundamentals of Speaker Recognition / B. Homayoon. – Springer, 2011. – P. 942.

4. *Huang X.* Spoken Language processing: a guide to theory, algorithm, and system development / X. Huang, A. Acero, H. Hon. – Prentice Hall PTR, 2001. – P. 936.

5. *Rabiner L.* Fundamentals of speech recognition / L. Rabiner, B.-H. Juang. – Prentice Hall PTR, 1993. – P. 507.

6. *Сергиенко А. Б.* Цифровая обработка сигналов / А. Б. Сергиенко – СПб. : Питер, 2002. – 608 с.

7. *Oppenheim Alan V.* Digital signal processing. Englewood Cliffs / Alan V. Oppenheim, Ronald W. Schafer – N.J: Prentice Hall, 1975. – pp. 548–554.

8. *Рабинер Л. Р.* Цифровая обработка речевых сигналов / Л. Р. Рабинер, Р. В. Шафер; пер. с анг. М. В. Назарова и Ю. Н. Прохорова. – М. : Радио и связь, 1981. – 496 с.

Запрягаев С. А. – д. физ.-мат. наук, профессор кафедры цифровых технологий, Воронежский государственный университет.
E-mail: zsa@cs.vsu.ru

Четкин А. С. – магистрант кафедры цифровых технологий, Воронежский государственный университет.
E-mail: andreychetkin@bk.ru

Zapryagaev S. A. – Dr. phys.-math. sciences, professor of the Department of digital technology, Voronezh State University.
E-mail: zsa@cs.vsu.ru

Chetkin A. S. – Graduate student of the Department of digital technology, Voronezh State University.
E-mail: andreychetkin@bk.ru