О сокращении перебора в словаре речевых команд в составе системы распознавания речи

И.Л. Мазуренко

Автор показал, что задачу распознавания речевых команд можно свести к решению математической задачи сравнения векторов различной длины методом динамического программирования. При этом количество элементарных шагов сравнений линейно зависит от числа команд в словаре и квадратично - от длины команды. При работе с большим словарем ($\cong 10,000$ слов) даже такие ограничения препятствуют распознаванию команд в реальное время. Поэтому для сокращения перебора предлагается сначала использовать быстрые алгоритмы предварительного распознавания. В данной статье приводится пример такого алгоритма.

1 Введение

Рассматривается задача распознавания речевых команд из заранее заданного словаря. Каждой команде однозначно сопоставляется ее текст и слово, составленное из символов звуков - транскрипция. Алгоритм имеет дело со звуковыми образами команд. Каждый звуковой образ представляет из себя звуковой сигнал - вектор значений звукового давления, измеренного со стандартной частотой. Обучающим материалом для алгоритма служат звуковые образы команд одного или нескольких дикторов (в зависимости от задачи).

Число звуковых образов команды и даже длин этих звуковых образов хотя и конечно, но так велико, что не представляется возможным использовать методику накопления эталонов команд с дальнейшим вычислением расстояния до ближайшего эталона. Содержательно, такое

разнообразие вариантов произнесения обусловлено неравномерным растяжением участков звуковых образов, соответствующих различным звукам.

Для сокращения размерности вместо множества звуковых сигналов $\bigcup_{N=N_0}^{N_1} I^N, I = \{i_0, i_0+1, \dots i_1\} \subset Z, i_0, i_1$ - константы, рассматривается множество описаний звуковых сигналов $\bigcup_{n=n_0}^{n_1} (J^M)^n, J = \{0, 1, \dots j_0 - 1\} \subset Z$, $n_0 = N_0/dT$, $n_1 = N_1/dT$, $|J^M| < |I^{dT}|$, где j_0, n_0, N_0, n_1, N_1, dT, M - константы. Предлагается в пространстве $\bigcup_{n=n_0}^{n_1} (J^M)^n$ определить расстояние между векторами $a \in (J^M)^{n_a}, b \in (J^M)^{n_b}$ адекватно восприятию речи человеком. Расстояние задается переборным алгоритмом, сложность которого может быть сокращена использованием метода динамического программирования.

Задача распознавания речи сводится к нахождения эталона, ближайшего к заданному звуковому сигналу. В качестве эталона команды рассматривается элемент пространства $\bigcup_{n=n_0}^{n_1} (J^M)^n$, являющийся усреднением звуковых образов этой команды, использованных для обучения.

Для успешного распознавания речевых команд необходимо решить следующие задачи:

правильно выбрать набор речевых команд;

выбрать компактный способ описания речевых сигналов;

построить эталоны, наиболее полно представляющие совокупность речевых образов каждой команды;

правильно определить расстояние между описаниями звуковых сигналов;

сократить перебор при нахождении ближайшего эталона.

Для решения последней задачи предлагается разбить звуки языка на классы (такие как "шипящие"звуки, "пауза", различные классы гласных звуков и т.п.) таким образом, что с высокой степенью надежности и независимо от диктора удается определить принадлежность участка сигнала к одному из этих классов. Звуковой сигнал можно закодировать последовательностью букв алфавита W, где W - это множество обозначений указанных классов.

Предлагается каждой речевой команде с сопоставить регулярное выражение $R(c)$, порождающее коды всех возможных звуковых образов ко-

манды c. В работе приведен алгоритм автоматического построения $R(c)$ по транскрипции команды c.

Установление близости звукового сигнала к эталону некоторой команды c сводится к проверке принадлежности кода сигнала к регулярному языку, порожденному эталонным кодом $R(c)$ этой команды.

Для нашего алгоритма этот подход позволяет априорно оценить попарную близость команд словаря и предсказать результаты распознавания этих команд.

2 Описание алгоритма распознавания речевых команд

Речевой командой c назовем совокупность $c = (l, z, \{s_i, i = 1, 2, ...n_c\}, e, R)$, где l-текст команды, z-ее транскрипция, $\{s_i, i = 1, 2, ...n_c\}$ - множество звуковых образов команды, e - эталон команды, R - эталонный код команды. Понятия текста, транскрипции, звукового образа, эталона и эталонного кода команды вводятся ниже.

Звуковым сигналом назовем целочисленный вектор $s = (x_1, \dots x_i, \dots x_N)$, где $x_i \in I = [-2^7, 2^7 - 1]$, являющийся последовательностью значений звукового давления в равноотстоящие друг от друга моменты времени.

Элементы x_i сигнала s будем называть *отсчетами*.

Будем считать, что *длина сигнала* N может принимать значения из множества $\{N_0, N_0 + 1, \dots N_1\} \in Z^+$. Множество всех звуковых сигналов обозначим через $S = \bigcup\limits_{N=N_0}^{N_1} I^N$.

Будем считать, что каждой речевой команде соответствует фиксированное множество звуковых сигналов - множество всех ее *звуковых образов*.

Разделим сигнал s на окна - векторы O_i длины T с шагом $dT, i = 1, ...n, n = [\frac{N-T}{dT}]+1, O_i = \left(x_{(i-1)\times dT}, x_{(i-1)\times dT+1}, \dots x_{(i-1)\times dT+j}, \dots x_{(i-1)\times dT+T-1}\right)$. i-е окно сигнала s будем обозначать через $s^{(i)}$.

Зафиксируем *функцию описания окна* $d : I^T \to J^M$, где M - некоторое натуральное число, множество $J = \{0, 1, \dots j_0-1\} \in Z^+$, такое что $|J^M| < |I^{dT}|$. Матрицу $D(s) = \|d(O_1)d(O_2)\dots d(O_p)\|$ назовем *описанием сигнала* s.

Подпоследовательность окон $\{O_j, O_{j+1}, \ldots O_k, 1 \leq j \leq n, 1 \leq k \leq n\}$, назовем *участком сигнала.*

В качестве функции описания окна сигнала будем рассматривать *спектр линейного предсказания*, рассчитываемый с помощью *коэффициентов линейного предсказания.*

Каждое окно сигнала $O_i = (x_0, x_1, \ldots x_{T-1})$ приблизим вектором $\hat{O}_i = (\hat{x}_0, \hat{x}_1, \ldots \hat{x}_{T-1}) \in \Re^T$, где

$$\hat{x}_n = \begin{cases} \sum\limits_{k=0}^{m} a_k \times k_{n-k}, & n \geq m \\ x_n, & n < m. \end{cases}$$

Коэффициенты $a_0, a_1, ...a_m$, выбираемые с помощью метода наименьших квадратов, то есть минимизирующие суммарную ошибку предсказания $\sum\limits_{j=0}^{T-1} |x_j - \hat{x}_j|^2$, называются *коэффициентами линейного предсказания.* Линейное предсказание окна длительностью T имеет размерность m, что позволяет значительно сократить размерность сигнала. В экспериментах автора для описания окна длиной в 512 отсчетов использовалось 20 коэффициентов линейного предсказания, причем без сколько-нибудь заметной потери полезной информации (критерием здесь является идентичность звукового сигнала s и его линейного предсказания $\hat{s}$ при прослушивании их человеком).

Использование вектора коэффициентов линейного предсказания в качестве функции описания окна неудобно для дальнейшего использования, поскольку метрика на таком множестве описаний сигналов оказывается весьма сложной. В приложениях используют модуль дискретного преобразования Фурье вектора коэффициентов линейного предсказания, который (это несложно объяснить с точки зрения некоторой простой физической модели речевого тракта человека) является огибающей *спектра* сигнала на окне анализа O_i (модуля дискретного преобразования Фурье окна анализа).

В качестве функции описания окна возьмем *спектр линейного предсказания* - функцию $G : I^T \to J^M$ модуля дискретного преобразования Фурье коэффициентов линейного предсказания:

$$G(O)_f = \left[\left| \sum_{j=0}^{m} a_j e^{-2\pi j \frac{f}{M} i} \right| \right],$$

где $f = 0, 1, ...M-1$ - номер частотной полосы, $a_j = a_j(O)$ - коэффициенты линейного предсказания, $O \in I^T$ - окно анализа.

Для расчета коэффициентов линейного предсказания надо минимизировать суммарную ошибку предсказания $\sum\limits_{j=m}^{T-1} \varepsilon_j^2$, где ошибка предсказания j-го отсчета $\varepsilon_j = |x_j - \hat{x}_j|$. Минимум суммарной ошибки достигается ([2], стр. 29) на решениях системы линейных уравнений

$$\sum_{k=0}^{m-1} a_k B(|k-j|) = B(j) \qquad (1),$$

где автокорреляция $B(k)$ выражается формулой $B(k) = \sum\limits_{i=0}^{T-k-1} x_i \times x_{i+k}$. Система (1) эффективно решается приближенным итерационным алгоритмом Дурбина [6]:

$$\left.\begin{array}{l} E(0) = B(0), E(s) = (1-k_s^2)E(s-1) \\ k_s = \dfrac{\left[B(s) + \sum\limits_{i=0}^{s-1} a_i^{(s-1)} B(|s-i|)\right]}{E(s-1)} \\ a_s^{(s)} = k_s, a_i^{(s)} = a_i^{(s-1)} + k_s a_{s-v}^{(s-1)}, v = 0, \ldots s-1 \end{array}\right\} s = 0, \ldots m$$

$$\text{решение } a_i = a_i^{(m)}, i = 0, \ldots m$$

Пусть O_1 и O_2 - окна некоторых звуковых сигналов. *Расстоянием* между O_1 и O_2 назовем величину $\rho_1(O_1, O_2) = \sum\limits_{i=0}^{M-1} |d(O_1)^{(i)} - d(O_2)^{(i)}|$, где $d(O_j)^{(i)}$ - i-й элемент описания окна O_j.

Растяжением звукового сигнала $s = (O_1, ...O_n)$ назовем последовательность окон

$$s[m_1, m_2, \ldots m_n] = (\underbrace{O_1, O_1, \ldots O_1}_{m_1}, \underbrace{O_2, O_2, \ldots O_2}_{m_2}, \ldots \underbrace{O_n, O_n, \ldots O_n}_{m_n}),$$

$$m_i > 0, i = 1, 2, \ldots n.$$

Длиной растяжения назовем количество входящих в него окон. Совокупность окон $\underbrace{O_i, O_i, \ldots O_i}_{m_i}$ в растяжении сигнала назовем i-й *группой окон.*

Через $< i >_s$ обозначим множество номеров всех окон в $s[m_1, m_2, ...m_n]$ из i-й группы: $< i >_s = \{m_1 + m_2 + \cdots + m_{i-1} + 1, m_1 + m_2 + \cdots + m_{i-1} + 2, \ldots m_1 + m_2 + \cdots + m_{i-1} + m_i\}$. Через $(j)_s$ обозначим группу окон, в которую входит окно с номером j, а через $[j]_s$ - номер этой группы.

Расстоянием между растяжениями s_1 и s_2 одинаковой длины n назовем величину $\rho_2(s_1, s_2) = \sum\limits_{i=1}^{n} \rho_1(O_i^1, O_i^2)$, где O_i^j- i-е окно растяжения s_j.

Расстоянием $\rho_3(s_1, s_2)$ между сигналами s_1 и s_2, состоящими из n_1 и n_2 окон соответственно, назовем величину

$$\min\{\rho_2(s_1[m_1^1, m_2^1, \ldots m_{n_1}^1], s_2[m_1^2, m_2^2, \ldots m_{n_2}^2]),$$
$$m_1^1 + m_2^1 + \cdots + m_{n_1}^1 = m_1^2 + m_2^2 + \cdots + m_{n_2}^2,$$
$$\forall i \ |(i)_{s_1}| > 1 \Rightarrow |(i)_{s_2}| = 1 \text{ и } |(i)_{s_2}| > 1 \Rightarrow |(i)_{s_1}| = 1\} \qquad (2),$$

где минимум берется по всем растяжениям сигналов s_1 и s_2 одинаковой длины, обладающим свойством $\forall i \ |(i)_{s_1}| > 1 \Rightarrow |(i)_{s_2}| = 1$ и $|(i)_{s_2}| > 1 \Rightarrow |(i)_{s_1}| = 1$ (3) (длина таких растяжений не превышает n_1+n_2, поэтому перебор конечен).

Такое определение расстояния между сигналами естественно, поскольку учитывает неравномерную длительность различных звуков в сигнале. Более того, растяжения $s_1[m_1^1, m_2^1, \ldots m_{n_1}^1]$ и $s_2[m_1^2, m_2^2, \ldots m_{n_2}^2]$ сигналов s_1 и s_2 соответственно, на которых достигается минимум при вычислении расстояния между s_1 и s_2, задают некоторое неоднозначное соответствие между окнами сигналов s_1 и s_2. Сопоставим номеру i окна O_i сигнала s_1 множество номеров $\bigcup\limits_{j \in < i >_{s_1}} [j]_{s_2}$ окон сигнала s_2. Обозначим это соответствие через $f : \{1 \ldots n_1\} \to 2^{\{1 \ldots n_2\}}$. Здесь через $2^{\{1 \ldots n_2\}}$ обозначено множество всех подмножеств множества $\{1 \ldots n_2\}$.

Для эффективного вычисления расстояния между звуковыми сигналами различной длины докажем следующее утверждение.

Пусть дана матрица $D = ||d_{ij}||$ размерности $m \times n$. *Путем* от элемента d_{pr} до элемента d_{kl}, $k \geq p$, $l \geq r$ назовем последовательность элементов $(d_{pr}, \ldots d_{ij}, d_{i'j'}, \ldots d_{kl})$, где $(i', j') = (i + 1, j)$ или $(i', j') = (i, j + 1)$ или $(i', j') = (i + 1, j + 1)$ (4); *весом пути* назовем сумму элементов в нем.

Утверждение 1. Пусть даны сигналы s_1 и s_2, состоящие из n_1 и n_2 окон соответственно, и матрица $D = ||d_{ij}||$ размерности $n_1 \times n_2$ попарных расстояний между окнами сигналов s_1 и s_2. Тогда

1) множество всех путей от d_{11} до $d_{n_1 n_2}$ находится во взаимнооднозначном соответствии со множеством всех растяжений одинаковой длины сигналов s_1 и s_2, удовлетворяющих условию (3).

2) вес любого пути от d_{11} до $d_{n_1 n_2}$ равен расстоянию между соответствующими ему растяжениями сигналов s_1 и s_2.

Доказательство.

Пусть $s_1[m_1^1, m_2^1, \dots m_{n_1}^1]$ и $s_2[m_1^2, m_2^2, \dots m_{n_2}^2]$ - растяжения сигналов s_1 и s_2 соответственно, причем $m_1^1 + m_2^1 + \dots + m_{n_1}^1 = m_1^2 + m_2^2 + \dots + m_{n_2}^2 = n$. Сопоставим этой паре растяжений путь $\left(d_{11}, \dots d_{[i]_{s_1}[i]_{s_2}}, \dots d_{n_1 n_2}\right)$. Свойство (4) пути выполнено в силу выполнения условия (3). Вес пути в точности совпадает с расстоянием между растяжениями сигналов s_1 и s_2.

Пусть теперь $\left(d_{11}, \dots d_{h(i)w(i)}, \dots d_{n_1 n_2}\right)$ - произвольный путь в матрице D. Сопоставим ему растяжения $\tilde{s}_1$ и $\tilde{s}_2$ с длиной, равной длине пути, такие что $\tilde{s}_1^{(i)} = s_1^{(h(i))}$, $\tilde{s}_2^{(i)} = s_2^{(w(i))}$. $\tilde{s}_1$ и $\tilde{s}_2$ - растяжения сигналов s_1 и s_2, так как функции $h(i)$ и $w(i)$ - номер строки и столбца i-го элемента пути - монотонны по i. Свойство (3) пары растяжений $\tilde{s}_1$ и $\tilde{s}_2$ выполняется в силу определения (4) пути. Вес пути совпадает с расстоянием между $\tilde{s}_1$ и $\tilde{s}_2$ по построению.

Утверждение доказано.

Следствие. Расстояние между сигналами s_1 и s_2 равно минимальному весу пути от d_{11} до $d_{n_1 n_2}$ в матрице попарных расстояний D.

Итак, нахождение расстояния между сигналами мы свели к нахождению пути с минимальным весом от левого верхнего до правого нижнего элемента прямоугольной матрицы. Эту задачу можно решить известным методом *динамического программирования* [5].

Действительно, для 1×1 матрицы решение тривиально. Если мы решаем задачу для $m \times n$ матрицы, зная решение для всех матриц $m' \times n'$, где $m' \leq m$ и $n' \leq n$, $(m', n') \neq (m, n)$, то можно выписать итерационную формулу для вычисления минимума: $d(m, n) = \min(d(m-1, n), d(m, n-1), d(m-1, n-1)) + d_{mn}$ (5). Здесь $d(m, n)$ - минимальный вес пути от d_{11} до d_{mn}.

Утверждение 2. Для нахождения расстояния между сигналами s_1 и s_2, состоящими из n_1 и n_2 окон соответственно, при условии, что известны попарные расстояния d_{ij} между окнами сигналов s_1 и s_2, необходимо не

более $3mn$ операций типа сложение / вычитание.

Доказательство.

Для каждой пары индексов окон $(i, j, 1 \leq i \leq m, 1 \leq j \leq m)$ вычисление расстояния $d(i, j)$ между усеченными сигналами $s_1^i = (s_1^{(1)}, s_1^{(2)}, \ldots s_1^{(i)})$ и $s_2^j = (s_2^{(1)}, s_2^{(2)}, \ldots s_2^{(j)})$ по формуле (5) динамического программирования, если уже вычислены все расстояния $\{d(k, l), k \leq i, l \leq j, (k, l) \neq (i, j)\}$, требует выполнения 2 операций сравнения и одной операции сложения. Всего упорядоченных пар $(i, j, 1 \leq i \leq m, 1 \leq j \leq m)$ усеченных сигналов ровно mn. Следовательно, для вычисления расстояния $d(s_1, s_2) = d(m, n)$ необходимо выполнение не более $(2 + 1)mn = 3mn$ операций типа сложение / вычитание.

Утверждение доказано.

Эталоном $e(c)$ речевой команды c, полученным по n звуковым образам $\{s_i(c), i = 1, \ldots n\}$ этой команды, назовем матрицу размерности $M \times \min\limits_{i=1\ldots n} |s^i(c)|$, определяемую следующим образом:

$$(e(c))_i = \frac{\sum\limits_{j=1\ldots n, j \neq j_0} \frac{\sum\limits_{k \in f^{j_0 j}(i)} d((s^j(c))^{(k)})}{|f^{j_0 j}(i)|} + d((s^{j_0}(c))^{(i)})}{n}$$

Здесь $j_0 = arg \min\limits_{i=1\ldots n} |s^i(c)|$, $s^{(i)}$ - i-е окно сигнала s, $d(.)$ - функция описания окна, $f^{ij}(.)$ - многозначное соответствие между окнами i-го и j-го звуковых образов команды c, $(e(c))_i$ - i-й столбец эталона.

Эталон, полученный по одному звуковому образу команды, очевидно, совпадает с описанием этого звукового образа. Эталон для n звуковых образов является усреднением их описаний с использованием соответствия $f(.)$ между их окнами.

i-й окном эталона e назовем его i-й столбец. Расстоянием между окном e_i эталона e и окном O_j звукового сигнала s назовем величину $\hat{\rho}_1(e_i, O_j) = \sum\limits_{k=0}^{M-1} |e_i^{(k)} - d(O_j)^{(k)}|$.

По аналогии определим расстояние $\hat{\rho}_2(.,.)$ между эталоном и звуковым сигналом, количество окон в котором равно длине эталона; растяжение эталона; расстояние между произвольным эталоном и произвольным звуковым сигналом.

Задача распознавания речи в терминах введенной модели будет сводится к нахождению минимума $\min\limits_{i=1\ldots n}(\hat{\rho}_3(e^i, s))$, где $\{e_i, i = 1\ldots n\}$ - множество эталонов, s - данный звуковой сигнал.

3 Сокращение перебора в словаре речевых команд

Алфавит L=*{а,б,в,г,д,. . . э,ю,я,-,'}* назовем *алфавитом русских букв*. Будем считать, что каждой речевой команде однозначно сопоставлено слово алфавита L - *текст команды*.

Алфавит Z = *{а, а', иа, иа', о, о', ио, ио', у, у', э, э', иэ, иэ', ы, ы', іы, іы', б, бь, в, вь, г, гь, д, дь, л, ль, м, мь, н, нь, р, рь, п, пь, к, кь, т, ть, с, сь, ф, фь, х, хь, ц, ч, ш, щ, й, ж, з, зь, _, ˜}* назовем *алфавитом русских звуков*, где _ означает паузу, а ˜ - звук-смычку, *а'* означает ударный а и т.д. Алфавит X = *{б, бь, в, вь, г, гь, д, дь, л, ль, м, мь, н, нь, р, рь, п, пь, к, кь, т, ть, с, сь, ф, фь, х, хь, ц, ч, ш, щ, й, ж, з, зь}* назовем *алфавитом шипящих звуков*. Будем считать, что каждому русскому звуку (как символу алфавита Z) соответствует некоторое множество звуковых сигналов, которые мы будем называть *звуковыми образами* звуков. Будем говорить, что некоторый участок сигнала соответствует русскому звуку z, если этот участок является звуковым образом звука z.

Пусть A=*{а'}*, O=*{о'}*, I=*{у'}*. Пусть также W={X, A, O, I, _, ˜}.

Известен один набор правил [3], осуществляющих перевод слов из алфавита русских букв в алфавит русских звуков (транскрибирование текста). Например, *“тся”* и *“ться”* заменяется на *“ца”*, *“лн”* на *“н”*, *“ого”* на *“ово”* и т.п. Нетрудно видеть, что эти правила можно описать в виде регулярной грамматики. Слово $T(z)$, получающееся из слова z в алфавите L, являющегося текстом некоторой речевой команды, в результате применения к нему правил этой грамматики, называется транскрипцией z. Транскрипцией команды назовем транскрипцию ее текста.

Регулярным выражением над алфавитом W называется формула, построенная из букв алфавита W и операций [ ], |, ·, * следующим образом:

- пустая формула $\emptyset$ - регулярное выражение
- a - регулярное выражение, если $a \in W$

- если X и Y - регулярные выражения, то $(X|Y)$ - регулярное выражение (где $(X|Y)$ это "X или Y")

- если X и Y - регулярные выражения, то XY - регулярное выражение (где XY это "конкатенация X и Y")

- если X - регулярное выражение, то X^* - регулярное выражение

Как следствие получаем, что если X - регулярное выражение, то $[X] = (\emptyset|X)$ - регулярное выражение.

Будем считать, что каждому русскому слову, как тексту некоторой речевой команды, соответствует некоторое множество звуковых сигналов, каждый из которых назовем звуковым образом этого слова.

Будем считать, что для большинства звуковых образов русского слова выполнены следующие условия:

Условие 1.

Любой звуковой сигнал, являющийся звуковым образом русского слова, можно разбить единственным образом на непересекающиеся участки двух видов: квазистационарные и переходные. *Квазистационарными* участками называются участки сигнала, соответствующие звукам транскрипции слова, а *переходными* - те участки сигнала, которые не могут быть отнесены к какому-либо отдельному звуку речи и соответствуют переходу от одного звука к другому.

б) Порядок следования квазистационарных участков сигнала тот же, что и порядок следования звуков в транскрипции слова, звуковым образом которого является сигнал.

Для некоторых звуков транскрипции может не быть соответствующего участка сигнала. Будем говорить, что в этом случае звук транскрипции в данном звуковом образе слова *пропускается*. Соседние звуки транскрипции пропускаться не могут.

в) Любой переходный участок по крайней мере в 2 раза короче каждого из соседних квазистационарных участков.

Условие 2.

Функция описания $d(.)$ выбрана таким образом, что для всех звуковых образов всех команд словаря описания окон из из квазистационарных участков, соответствующих каждому из шести классов звуков X, A, O, I, $\{\_\}$, $\{\tilde{}\}$, принадлежат непересекающимся подмножествам множества J^M. Описания остальных окон из квазистационарных участков принадлежат объединению подмножеств множества J^M, соответствую-

щих классам A, O и I. Обозначим эти подмножества через $Q(X)$, $Q(A)$, $Q(O)$, $Q(I)$, $Q(\_)$ и $Q(\tilde{})$ соответственно.

Условие 2 для простоты сформулировано в таком виде. Реально спектр линейного предсказания гарантирует с большой вероятностью восстановление класса из W по описанию окна.

Для каждого звукового сигнала s произведем классификацию всех окон этого сигнала на классы из W (если описание окна принадлежит одному из множеств $Q(X)$, $Q(A)$, $Q(O)$, $Q(I)$, $Q(\_)$ и $Q(\tilde{})$, считаем, что окно принадлежит классу X, $A, O, I, \{\_\}, \{\tilde{}\}$ соответственно). Полученное слово w в алфавите W назовем *длинным кодом сигнала*. Из условия 1в следует, что в слове w можно отличить квазистационарные участки от переходных (по их длине). Значит, можно построить автомат, обрабатывающий длинные коды сигналов, на выходе которого будут слова в алфавите W (назовем их *кодами звуковых сигналов*), последовательность символов в которых будет соответствовать последовательности квазистационарных участков сигнала. Код звукового сигнала s обозначим через $K(s)$.

Будем считать, что для каждого слова заранее известны звуки из его транскрипции, которые могут пропускаться во всех звуковых образах этого слова (это чаще всего глухие согласные в конце слова, взрывные согласные п,к,т и т.п.). Под эталонным кодом команды c будем понимать регулярное выражение $R(c)$, формула которого получается из транскрипции $T(c)$ команды c так:

1) Если звук из транскрипции может пропускаться, заключаем этот звук в квадратные скобки.

2) Если звук принадлежит одному из шести введенных классов звуков, заменяем его на символ этого класса.

3) Остальные звуки заменяем на регулярное выражение $(A|O|I)$.

Например, эталоном слова “четыре”(транскрипция *чи_ ты'ри*) будет регулярное выражение $X[(O|I|A)]\_[X]I[X][(A|O|I)]$, если известно, что оба безударных звука *и*, звуки *т* и *р* могут пропускаться при произнесении.

Пусть L_1 и L_2 - *регулярные языки* [4], порожденные выражениями $R(c_1)$ и $R(c_2)$, где c_1 и c_2 $O \in A^*$. Если $L_1 \cap L_2 = \emptyset$, то будем говорить, что - c_1 и c_2 *совместны*. Будем также говорить, что слово w в алфавите W *удовлетворяет* эталону R, если оно принадлежит языку, порожденному

R (обозначение $w \in R$).

Следующая теорема содержательно означает, что если некоторый звуковой сигнал s правильно распознается полнопереборным алгоритмом, то в перебор можно включать только те команды, эталонным кодам которых удовлетворяет код сигнала s.

Теорема. Пусть s - звуковой сигнал, удовлетворяющий условиям 1 и 2, $(e_1, e_2, \ldots e_n)$ - эталоны речевых команд $(c_1, c_2, \ldots c_n)$. Пусть сигнал s является звуковым образом команды c_j.

Если $c_j = arg\min\{\hat{\rho}_3(s, e_i), i = 1, \ldots n\}$, то

$c_j = arg\min\{\hat{\rho}_3(s, e_i), K(s) \in R(\mathrm{c}_i), i = 1, \ldots n\}$

Доказательство.

По построению код любого образа команды c_i удовлетворяет ее эталонному коду $R(c_i), i = 1, \ldots n$. Тогда $c_j \in \{\hat{\rho}_3(s, e_i),\ K(s) \in R(c_i), i = 1, \ldots n\}$, откуда $arg\min\{\hat{\rho}_3(s, e_i), i = 1, \ldots n\} = arg\min\{\hat{\rho}_3(s, e_i), K(s) \in R(c_i), i = 1, \ldots n\}$.

Теорема доказана.

4 Заключение

На основе описанной модели автором была разработана компьютерная распознающая система Voice Commander, позволяющая с предварительной настройкой на диктора продемонстрировать работу алгоритма распознавания на произвольном словаре объемом до 100 слов.

Для написания программы использовался язык программирования Visual C++ 1.5 (операционная система Windows 3.1). Программа работает в реальном времени на компьютере типа IBM PC с процессором Pentium 100 МГц. Результаты распознавания зависят от словаря команд, средняя ошибка распознавания для словаря из 10 команд составляет 3%.

Работа выполнена на кафедре Математической теории интеллектуальных систем механико-математического факультета Московского Государственного Университета им. М. В. Ломоносова.

Автор выражает благодарность своему научному руководителю кандидату физико-математических наук Бабину Дмитрию Николаевичу за помощь в написании статьи и разработке компьютерной распознающей системы.

Список литературы

[1] Мазуренко И. Л. *Одна модель распознавания речи.* В сб.: Компьютерные аспекты в научных исследованиях и учебном процессе (по материалам научной конференции МГУ "Ломоносовские чтения - 96") М.: Издательство Московского университета, 1996.

[2] Винцюк Т. К. *Анализ и распознавание речевых сигналов.* Киев, 1985.

[3] Богданова Н. В. , Бондарко Л. В. , Ипатов Я. В. , Коваль С. Л. , Овчаренко С. Б. , Панова-Яблошникова И. С. , Степанова С. Б. *Автоматический транскриптор в системах распознавания, синтеза и обучения.* // АРСО-15

[4] Кудрявцев В. Б. , Алешин А. В. , Подколзин А. С. *Введение в теорию автоматов.* - М. : Наука, 1985.

[5] Беллман Р. *Динамическое программирование.* - М. : Издательство иностранной литературы, 1960.

[6] Андерсен Т. У. *Введение в многомерный статистический анализ.* - М. : 1960.