

Библ. 6
А-43

КАУНАССКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ

А. Ю. М О Т У З А

**Исследование процессов выбора
наилучших вариантов опознающих
устройств**

05.253 Приборы и устройства автоматики
и телемеханики

А в т о р е ф е р а т
диссертации на соискание ученой степени
кандидата технических наук

КАУНАС - 1969

Одной из актуальных проблем на современном этапе развития науки и техники является проблема опознавания образов. Эта проблема охватывает такие важные научные и технические задачи, как создание устройств для автоматического чтения машинописных и рукописных текстов, автоматов опознавания речевых сигналов, устройств для для медицинской диагностики болезней, систем для интерпретации данных геологической разведки.

Важной задачей опознавания образов в настоящее время является создание эффективных по выбранным критериям опознающих устройств. При решении этой задачи часто приходится рассматривать большое число вариантов опознающего устройства с целью выбора наилучшего по заданному критерию варианта. Такими вариантами могут быть, например, различные системы признаков [1] и различные режимы работы читающего устройства [2].

В ряде случаев основным критерием качества Q систем опознавания является вероятность ошибки опознавания $P_{ош}$ (или вероятность отказа от опознавания $P_{от}$). Тогда наилучшим из рассматриваемых вариантов опознающей системы является тот вариант, для которого Q является минимальным.

Значения вероятностей $P_{ош}$ ($P_{от}$) точно неизвестны. Практически они оцениваются по результатам исследований выборок опознаваемых объектов. При конечных объемах этих выборок оценки показателей качества вариантов опознающей системы являются случайными величинами. Разброс этих оценок может привести к тому, что вариант, выбранный по наименьшей оценке показателя качества Q не всегда будет наилучшим из рассматриваемых вариантов опознающей системы, что приведет к уменьшению эффективности использования опознающей системы.

Вероятность ошибки выбора наилучшего варианта можно сколько угодно уменьшать увеличением объемов выборок опознаваемых объектов. Но увеличение этих объемов связано с затратами. Например, в случае создания опознающих устройств автоматического чтения машинописных или рукописных текстов, увеличение объема выборки связано с проведением долговременных испытаний, в случае опознающих систем для медицинской диагностики - с про-

ведением сложных медицинских исследований. В связи с этим возникает следующий принципиально важный вопрос. Каким должен быть наименьший объем выборки опознаваемых объектов, по которым оценивается качество выбираемых вариантов, чтобы он гарантировал заданную степень достоверности выбора наилучшего варианта? Практически выбор считается достоверным, если вероятность P выбора наилучшего варианта (вероятность правильного выбора) больше или равна некоторой заданной заранее величине P^* .

В реферируемой работе рассматривается задача выбора наилучшего варианта опознающего устройства из заданного числа S вариантов в случае, когда для оценки качества этих вариантов производится испытание опознающего устройства. Во время каждого испытания опознается выбранное заранее число N опознаваемых объектов и подсчитывается число случаев ошибочного опознавания (отказов от опознавания). Тогда результаты испытаний каждого i -ого ($i = 1, \dots, S$) варианта можно рассматривать как биномиально распределенную случайную величину X_i . Эта случайная величина определяется вероятностью ошибки опознавания для i -ого варианта $P_{ош}^{(i)}$ (вероятностью отказа от опознавания $P_{от}^{(i)}$) и числом N опознаваемых объектов, используемых для одного испытания. При малых значениях вероятностей $P_{ош}^{(i)}$ ($P_{от}^{(i)}$) распределение результатов испытаний X_i с достаточной для практических расчетов точностью аппроксимируется распределением Пуассона с параметром $Q_i = P_{от}^{(i)} \cdot N$ ($Q_i = P_{ош}^{(i)} \cdot N$) названным показателем качества i -ого варианта опознающего устройства. Тогда задача выбора наилучшего варианта сводится к задаче выбора той случайной величины X_m из S распределенных по закону Пуассона случайных величин, параметр распределения которой Q_m является наименьшим.

В ряде практических случаев задача выбора наилучшего варианта считается решенной, если выбран тот j^* -ый вариант $j^* \in \{1, \dots, S\}$ показатель качества которого мало (не больше чем на заданную величину) отличается от минимального показателя качества Q_m , т.е. выбор считается правильным, если выбирается тот вариант, для которого выполнено неравенство:

$$Q_{j^*} \leq Q_m (1 + \Delta), \quad (1)$$

где Δ - заданная заранее константа. Если, например, выбрано $\Delta = 0,25$, то показатель качества выбираемого варианта может превышать минимальный показатель качества не больше, чем на 25%.

В диссертации предлагаются и исследуются процедуры, гарантирующие заданную вероятность P^* выбора варианта опознающей системы, удовлетворяющего условию (1).

Работа состоит из введения, семи глав и трех приложений. В первой главе формулируются задачи, дается краткий обзор ра-

бот по проблеме выбора, выясняются нерешенные задачи данной области.

x
x x

Во второй главе предлагается и исследуется методика определения числа испытаний N^* , при котором вероятность P выбора варианта, удовлетворяющего условию (1) равна или превышает заданное значение P^* . Рассматривается случай, когда числа испытаний для каждого i -ого варианта равны между собой ($N_1 = \dots = N_s = N$) и когда величины X_1, \dots, X_s независимы. При указанных условиях рассматривается процедура (названная одновыборочной процедурой) определяемая так:

Выбирается тот j -ый вариант, для которого оценка качества

$$y_j = \sum_{k=1}^N x_{jk}, \quad (2)$$

где x_{jk} - реализация случайной величины X_j , является наименьшей. В случае, когда для нескольких вариантов наименьшие оценки равны между собой, из этих вариантов случайно выбирается один вариант.

В случае, когда результаты испытаний X_1, \dots, X_s вариантов опознающей системы распределены по закону Пуассона, получено следующее выражение для требуемого числа испытаний N^* :

$$N^* = \left[\frac{I^*}{Q} \right]^* + 1, \quad (3)$$

где Q - заданная заранее нижняя граница наименьшего показателя качества Q_m ,

$[\eta]^*$ - целая часть числа η .

I^* - параметр - нами названный приведенным объемом выборки.

Показано, что приведенный объем выборки I^* равен положительному (> 0) решению уравнения

$$\sum_{n=0}^{\infty} f(n|I) \sum_{i=0}^{s-1} \frac{C_{s-1}^i}{i+1} f^i [n|I(1+\Delta)] \{1 - F[n|I(1+\Delta)]\}^{s-1-i} = P^* \quad (4)$$

относительно неизвестного I ,

где

$$f(n|\eta) = \frac{e^{-\eta} \eta^n}{n!}, \quad F(n|\eta) = e^{-\eta} \sum_{j=0}^n \frac{\eta^j}{j!}.$$

Положительное решение уравнения (4) всегда существует при $\rho^* > 1/s$, $\Delta > 0$, $s > 1$.

Для наиболее часто встречающихся значений вероятности ρ^* и параметров Δ и s приведенный объем выборки J^* был табулирован. Часть результатов приводится в таблице 1

Таблица 1
Величина J^* в зависимости от ρ^* , s и Δ .

| ρ^* | $s=2$ | | $s=4$ | | $s=8$ | |
|----------|--------------|--------------|--------------|--------------|--------------|--------------|
| | $\Delta=1,0$ | $\Delta=0,5$ | $\Delta=1,0$ | $\Delta=0,5$ | $\Delta=1,0$ | $\Delta=0,5$ |
| 0,80 | 2,16 | 7,13 | 5,49 | 18,2 | 8,40 | 27,9 |
| 0,90 | 4,86 | 16,3 | 9,02 | 30,2 | 12,4 | 41,4 |
| 0,95 | 8,01 | 26,8 | 12,6 | 42,3 | 16,3 | 54,5 |

Пусть, например, $\rho^* = 0,9$, $Q = 0,01$, $\Delta = 0,5$, $s = 4$. Тогда воспользовавшись таблицей 1 и формулой (3) находим $N^* = 3021$.

Показано, что в случае, когда величины X_1, \dots, X_s распределены биномально, требуемое число испытаний N^* , вычисленное с использованием пуассоновской аппроксимации является завышенным. С целью количественной оценки этой завышенности были подсчитаны точные и приближенные значения величины N^* .

Анализ численных результатов показал, что при пуассоновской аппроксимации биномиального распределения величины X_1, \dots, X_s и при $Q < 0,1$ требуемое число испытаний является завышенным не больше, чем на 5%. Такая точность является достаточной для практических расчетов.

Предложенная методика была исследована экспериментально с использованием читающего устройства „РУТА-701“ [3]. Условия проведения экспериментов на этом устройстве и краткое описание самого устройства приводится в приложении 1 диссертации.

Алгоритм опознавания исследуемого устройства описывается линейной функцией

$$R_k = \sum_{j=1}^o w_{jk} x_j + \sum_{j=1}^o \bar{w}_{jk} (1 - x_j),$$

$$k = 1, \dots, K, (x_j, w_{jk}, \bar{w}_{jk} = 0; 1),$$

где R_k — коэффициент подобия опознаваемого объекта с k -ым образом, x_j — признаки, w_{jk} , \bar{w}_{jk} — веса, K — число классов опознаваемых объектов, o — число признаков. В качестве признаков x_j используются значения сигналов, соответствующих j -ым

элементам считываемого знака ($x_j = 1$ для „черного“ элемента изображения знака и $x_j = 0$ для „белого“).

Опознаваемый объект относится к тому классу (если принимается решение) для которого коэффициент подобия максимальный. Решение принимается в случае, когда максимальный коэффициент подобия и разница между максимальным коэффициентом и ближайшим к нему превышают заданные пороги $a_{пор}$ и $b_{пор}$, соответственно. Если хотя бы одно из этих условий не выполнено устройство отказывается от решения.

От значений порогов $a_{пор}$ и $b_{пор}$ зависят значения вероятностей ошибочных решений и отказов от решений. Поэтому при проведении исследований устройства требовалось выбрать такие значения этих порогов, при которых указанные вероятности были бы наименьшими.

Для упрощения экспериментов критерием качества Q режимов было выбрано суммарное среднее число ошибочно опознанных или неопознанных (случаев отказов от решений) объектов из числа объектов, используемых в одном испытании. Для каждого испытания использовалось по 580 объектов — машинописных арабских цифр, напечатанных на одном бланке.

Для исследования выбирались те режимы работы устройства, для которых из предыдущих экспериментов были известны значения принятого критерия качества режимов Q — среднего числа случаев ошибочных решений и отказов от решений. Для этих режимов теоретически подсчитывались значения вероятности P выбора наилучшего варианта при различных значениях числа испытаний N и числа вариантов s . Вероятность P так же была оценена экспериментально с использованием данных, полученных на читающем устройстве. Часть результатов расчетов и экспериментов — вероятности P , оценки этих вероятностей \tilde{P} , нижние $P^{(H)}$ и верхние $P^{(B)}$ доверительные границы величины \tilde{P} , подсчитанные при уровне значимости 5% приводятся в таблице 2.

Таблица 2
Экспериментальные результаты ($s = 3$, $Q_1 = 1,10$; $Q_2 = 1,60$; $Q_3 = 1,75$)

| N | 2 | 4 | 6 | 8 | 10 |
|-------------|-------|-------|-------|-------|-------|
| P | 0,547 | 0,642 | 0,709 | 0,760 | 0,800 |
| \tilde{P} | 0,48 | 0,61 | 0,68 | 0,78 | 0,81 |
| $P^{(H)}$ | 0,46 | 0,56 | 0,63 | 0,69 | 0,73 |
| $P^{(B)}$ | 0,63 | 0,72 | 0,78 | 0,83 | 0,86 |

Проведенные экспериментальные исследования показали хорошее для практических расчетов согласие теоретических результатов с экспериментальными данными.

x x
x x

Решение уравнения (4) связано с трудоемкими вычислениями, которые могут быть выполнены только с использованием быстродействующих цифровых вычислительных машин. Поэтому в третьей главе рассматриваются способы приближенного расчета приведенного объема выборки I^* . Отдельно рассматриваются случаи $s = 2$ и $s > 2$.

В случае $s = 2$, с использованием формулы, устанавливающей связь между нецентральным распределением χ^2 и распределением Пуассона [4] показано, что решения уравнений

$$F_{\chi^2} [2I | 2I(1 + \Delta), 2] = 1 - P^*, \quad (6)$$

$$F_{\chi^2} [2I(1 + \Delta) | 2I, 2] = P^*$$

относительно неизвестного I являются нижней \underline{I}^* и верхней \bar{I}^* границами приведенного объема выборки I^* . Здесь $F_{\chi^2} [q | \lambda, 2]$ - функция нецентрального распределения χ^2 с параметром нецентральности λ и двумя степенями свободы.

Для практического расчета нижней \underline{I}^* и верхней \bar{I}^* границ приведенного объема выборки I^* получены следующие выражения, основанные на трансформации Абдел-Ату [5] нецентрального распределения χ^2 :

$$\underline{I}^* = \frac{I(\theta_1)}{1 + \alpha}; \quad \bar{I}^* = I(\theta_2),$$

где

$$\alpha = \sqrt[3]{1 + \Delta} - 1,$$

$I(\theta)$ - функция, обратная функции

$$\theta = 3(1 + \Delta) \frac{\sqrt[3]{\frac{I}{1 + I}}}{\sqrt{1 + 2I}},$$

$$\theta_1 = k_1 - \frac{\alpha - 1}{\alpha} \frac{\varepsilon^2(k_1)}{2\varepsilon(k_1) - \varepsilon[k_1 - \varepsilon(k_1)]},$$

$$\theta_2 = k_2 + \frac{\varepsilon^2(k_2)}{2\varepsilon(k_2) + \varepsilon[k_2 + \varepsilon(k_2)]},$$

$$k_1 = k_2(\alpha + 1),$$

$$k_2 = \frac{\Phi^{-1}(P^*)}{\alpha},$$

$\Phi^{-1}(P^*)$ - функция обратная функции нормального распределения вероятностей,

$$\varepsilon(\theta) = \frac{9[1 + I(\theta)]^2 - 1 - 2I(\theta)}{3\sqrt{1 + 2I(\theta)} [1 + I(\theta)]} - \theta.$$

Для приближенного расчета приведенного объема выборки предложено использовать следующее выражение:

$$I^* \approx I_c = \frac{\underline{I}^* + \bar{I}^*}{2}. \quad (7)$$

С целью оценки точности этой аппроксимации были подсчитаны значения величин \underline{I}^* и \bar{I}^* при различных значениях величин P^* и Δ .

Анализ результатов проведенных расчетов показал достаточную для практических целей точность предлагаемой аппроксимации. Для облегчения практического применения полученных выражений функции $\varepsilon(\theta)$ и $I(\theta)$ были табулированы.

В случае $s > 2$ для определения приведенного объема выборки I^* предложено использовать результаты, полученные при $s = 2$. Для этого доказано следующее:

Утверждение. Приведенный объем выборки $I^{*(1)}$ подсчитанный так, чтобы в случае двух вариантов ошибка выбора $1 - P$ не превышает $(1 - P^*) / (s - 1)$ гарантирует заданную вероятность правильного выбора P^* при числе вариантов равным s ($s > 2$).

Из этого утверждения следует, что для нахождения приведенного объема выборки в случае $s > 2$ можно воспользоваться методикой, полученной при $s = 2$. При этом величина P^* должна быть заменена величиной $(s - 2 + P^*) / (s - 1)$.

Приведенный объем выборки $I^{*(1)}$ при $s > 2$, подсчитанный с использованием утверждения, является завышенным. Для оценки этой завышенности были подсчитаны значения величин \underline{I}^* и \bar{I}^* при различных значениях параметров s , Δ и P .

Путем исследования результатов расчетов было найдено, что при $P^* \geq 0,9$ и $s \leq 8$ увеличение приведенного объема выборки I^* вызываемое аппроксимацией, основанной на выше указанном утверждении не превышает 20%.

x x
x x

Одновыборочная процедура, рассмотренная во второй и третьей главах, имеет следующий недостаток. Она полностью использует информацию, которая находится в результатах испытаний опознающей системы. Это приводит к возрастанию числа испытаний. По-

этому в четвертой главе исследуются последовательные процедуры выбора, при использовании которых число испытаний не определяется заранее, а последовательно увеличивается. При этом момент прекращения процесса выбора зависит от результатов испытаний предшествующих указанному моменту.

Процедуры выбора наилучшего варианта должны быть эффективными по критерию числа испытаний и удобными для практического использования. С целью определения процедуры, удовлетворяющей вышеуказанным условиям, были рассмотрены следующие последовательные процедуры:

1. Последовательная процедура, основанная на вычислении апостериорных вероятностей величин Q_1, \dots, Q_s (названная процедурой A_1);

2. Две последовательные процедуры (названные процедурами A_2 и A_3), полученные с использованием леммы Вальда, устанавливающей связь между двумя распределениями последовательности случайных величин ([6] стр. 188).

Процедура A_1 определяется так: процесс выбора разбивается на шаги. На каждом l -ом шаге ($l = 1, 2, \dots$) выполняются следующие операции:

- Определяются оценки качества вариантов y_1, \dots, y_s ,
- Проверяется неравенство

$$\sum_{\substack{j=1 \\ j \neq m}}^s (\alpha^*)^{y_j - y_m} \leq \frac{1 - \rho^*}{\rho^*}, \quad (8)$$

где

$$\alpha^* = \begin{cases} \frac{1}{1 + \Delta} - \frac{\Delta Q}{(1 + \Delta)(1 - Q)} & \text{в случае биномиального} \\ & \text{распределения } X_1, \dots, X_s, \\ \frac{1}{1 + \Delta} & \text{в случае пуассоновского} \\ & \text{распределения } X_1, \dots, X_s. \end{cases}$$

Если неравенство (8) выполнено, последовательный процесс прекращается выбором m -ого варианта;

в) В случае, когда неравенство (*) не выполнено, увеличивается число испытаний $N = N_1 = \dots = N_s$ и выполняются операции, указанные в пунктах а и б.

Процедура A_2 отличается от процедуры A_1 условием выбора m -ого варианта. Это условие для процедуры A_2 определяется так:

$$\sum_{\substack{k=1 \\ k \neq j}}^s (\alpha^*)^{y_k - y_j} > \frac{s-1}{1 - \rho^*} \quad (9)$$

для всех $j = 1, \dots, s; j \neq m$.

В случае использования процедуры A_3 на каждом l -ом шаге процесса выбора выбрасываются из дальнейшего рассмотрения те J -ые варианты, для которых

$$y_j > \frac{\ln \frac{1 - \rho^*}{s-1}}{\ln \alpha^*} + \min_{i \in J_{l-1}} y_i, \quad (10)$$

где J_{l-1} - множество индексов вариантов, не выброшенных на первом, ..., $l-1$ -ом шагах процесса выбора. Процесс выбора продолжается до тех пор, пока не остается только один невыброшенный вариант.

Эффективность (по критерию среднего числа испытаний $E = E(\sum_{i=1}^s N_i)$) использования процедур A_1, A_2 и A_3 исследовалась методом математического моделирования и экспериментально. Для проведения этих исследований были составлены программы, реализующие процедуры A_1, A_2 и A_3 и моделирующие работу вариантов опознающего устройства. При помощи этих программ были получены оценки \bar{E} среднего числа испытаний E и оценки $\bar{\sigma}$, среднеквадратических отклонений величин E .

Рассматривался случай, когда оценки качества вариантов распределены по закону Пуассона и когда Q_1, \dots, Q_s имеют „наименее благоприятную“ расстановку: т.е. $Q_j = Q_m (1 + \Delta)$ для всех $j = 1, \dots, s, j \neq m$.

Путем исследования результатов математического моделирования было установлено, что при использовании процедур A_1 и A_3 вместо одновыборочной процедуры среднее число испытаний можно уменьшить до двух раз. Эффективность использования этих процедур возрастает с увеличением числа вариантов и с уменьшением величины Δ , определяющей расстояние между показателями качества наилучшего варианта и других вариантов.

Процедура A_2 во всех случаях (кроме $s = 2$, где она совпадает с процедурой A_3) требует большего среднего числа испытаний, чем процедуры A_1 и A_3 или даже чем одновыборочная процедура.

В случае, когда значения показателей качества Q_1, \dots, Q_s отличаются от „наименее благоприятных“, эффективность использования процедур A_1, A_2, A_3 была исследована экспериментально с использованием читающего устройства „РУТА-701“ [3]. Во время проведения этих экспериментов выбирался один из семи заданных заранее режимов работы устройства. Показателями качества Q режимов были выбраны средние числа неопознанных знаков из заданного числа (580) знаков напечатанных на одном бланке.

Выбор варианта считается правильным, если выбирался j -ый режим работы устройства, удовлетворяющий условию (1).

Эксперименты выбора проводились для всех трех процедур при различных значениях величин P^* и Δ . Часть результатов этих экспериментов - оценки среднего числа испытаний \bar{E} и оценки среднеквадратических отклонений $\tilde{\sigma}$ величины \bar{E} приводятся в таблице 3.

Таблица 3

Величины \bar{E} и $\tilde{\sigma}$ в зависимости от Δ ($s=7, Q_1=1,00, Q_2=1,50, Q_3=1,80, Q_4=Q_5=1,75, Q_6=3,00, Q_7=7,00, P^*=0,8$)

| Процедура | A_1 | | A_2 | | A_3 | |
|------------------|-------|-----|-------|-----|-------|-----|
| Δ | 1,0 | 0,5 | 1,0 | 0,5 | 1,0 | 0,5 |
| \bar{E} | 86 | 89 | 107 | 140 | 38 | 72 |
| $\tilde{\sigma}$ | 5,4 | 5,6 | 7,0 | 0,0 | 1,6 | 1,4 |

Анализ результатов проведенных экспериментов показал, что в случае, когда значения показателей качества Q_1, \dots, Q_s отличаются от „наименее благоприятных“, процедура A_3 требует значительно меньшего среднего числа испытаний, чем процедуры A_1 или A_2 . Кроме того эта процедура является наиболее удобной для практического использования из всех процедур A_1, A_2 и A_3 .

В случае, когда число вариантов равно 2 и когда оценки y_1 и y_2 распределены биномиально, была исследована зависимость основных характеристик последовательных процедур - вероятности P выбора наилучшего варианта и среднего числа испытаний E от степени зависимости между величинами X_1 и X_2 . С этой целью были получены зависимости величины X_1 и X_2 от коэффициента корреляции ρ между результатами испытаний X_1 и X_2 .

Путем исследования полученных выражений установлено:

1. Вероятность выбора наилучшего варианта P и среднее число испытаний E являются возрастающими функциями от коэффициента корреляции ρ .

2. $\lim_{\rho \rightarrow 1} P \rightarrow 1;$

3. В случае, когда выбрано $Q = 0$, последовательные процедуры A_1, A_2 и A_3 гарантируют заданную вероятность P^* выбора наилучшего варианта опознающей системы при любом значении коэффициента корреляции ρ .

4. В случае, когда $|Q_1 - Q_2| > 0$ корреляция между случайными величинами X_1 и X_2 может увеличить среднее число испытаний не больше чем $1/(2P^* - 1)$ раз.

5. Если $Q_1 - Q_2 = 0, \lim_{\rho \rightarrow 1} E \rightarrow \infty.$

Процедуры выбора, рассмотренные в четвертой главе, получены при условии, что максимальное возможное число испытаний неограничено сверху. Практически это условие никогда точно не выполняется. Поэтому в пятой главе исследуется усеченная последовательная процедура A_y , полученная из процедуры A_3 усечением процесса выбора заранее заданной величиной N_0 . Процедура A_y определяется так:

Процесс выбора разбивается на шаги, максимальное возможное число которых равно N_0 . На каждом N -ом шаге выполняются следующие операции:

а) на единицу увеличивается число испытаний N и подсчитываются оценки качества вариантов, не исключенных из дальнейшего исследования на 1-ом, ..., N -ом шагах процесса выбора.

б) Исключаются из дальнейшего рассмотрения те j -ые варианты для которых

$$y_j > a + \min_{i \in J_{N-1}} y_i \quad \text{при } N < N_0$$

или

$$y_j > \min_{i \in J_{N-1}} y_i \quad \text{при } N = N_0,$$

где a_1 - заданное заранее натуральное число, J_{N-1} - множество индексов вариантов, не исключенных из дальнейшего рассмотрения на первом, ..., $N-1$ -ом шагах процесса выбора.

Процесс выбора продолжается до тех пор, пока не остается только один не исключенный вариант или когда число испытаний не достигает N_0 . В случае, когда для нескольких вариантов наименьшие оценки y_i равны между собой и когда $N = N_0$, из этих вариантов случайно выбирается один.

Показано, что предлагаемая процедура A_y в случае биномиального или пуассоновского распределений величин X_1, \dots, X_s описывается цепью Маркова [7] с конечным числом состояний. При $s=2$ получены точные рекуррентные выражения для вероятности выбора наилучшего варианта $P_N = P(N_0, a_1)$ и для среднего числа испытаний $E_N = E(N_0, a_1)$ определяющие зависимость указанных величин от N_0 и a_1 .

При $s > 2$ для определения вероятности $P(N_n, a_1)$ и среднего числа испытаний $E(N_0, a_1)$ предложены приближенные выражения. Точность этих выражений исследовалась методом математического моделирования. Анализ результатов математического моделирования показал высокую точность полученных приближенных выражений.

Полученные зависимости вероятности правильного выбора позволяют определить такие значения параметра a_1 , усеченной последовательной процедуры A_s , при которых она гарантирует заданную вероятность P^* выбора наилучшего варианта.

Поскольку вероятность P_{N_0} и среднее число испытаний E_{N_0} — возрастающие функции от a_1 , из всех значений величины a_1 , удовлетворяющих условию $P_{N_0} \geq P^*$, должно быть выбрано наименьшее значение a_1^* . Это значение может быть найдено с использованием полученного выражения для $P(N_0, a_1)$. Для этой цели также могут быть использованы таблицы значений P_{N_0} , подсчитанные при различных значениях величины N_0, a_1, Q_m и Δ . Такие таблицы приводятся в реферируемой работе. Часть результатов расчетов — значения величины P_{N_0} при $Q_m = 1; \Delta = 0,5$ приводится в таблице 4.

Таблица 4

| $N_0 \backslash a_1$ | 4 | 5 | 6 | 7 | 8 | 9 |
|----------------------|-------|-------|-------|-------|-------|-------|
| 20 | 0,889 | 0,905 | 0,914 | 0,919 | 0,921 | 0,921 |
| 40 | 0,902 | 0,931 | 0,949 | 0,961 | 0,986 | 0,972 |

Пусть, например, $s = 2, \Delta = 0,5, Q_m = 1, N_0 = 40, P^* = 0,95$ тогда из таблицы 4 находим, что наименьшее значение a_1^* величины a_1 , при которой $P_{N_0} \geq P^*$ равна 7.

Эффективность использования предлагаемой процедуры была исследована при помощи так называемого выигрыша в среднем числе испытаний [8]. Выигрыш B вычисляется по формуле:

$$B = \left(1 - \frac{E}{N^*}\right) \cdot 100\% \quad (11)$$

С использованием выражения (11), а также методики, приведенной во второй главе реферируемой работы, было установлено, что выигрыш B имеет тенденцию возрастать при увеличении вероятности P^* и параметра N_0 . В случае, когда $P^* \geq 0,95$ выигрыш B достигает 50%. При малых значениях вероятности P^* ($P^* \leq 0,8$) выигрыш B может быть даже отрицательным. Поэтому последовательные процедуры целесообразно использовать только при больших значениях вероятности P^* ($P^* \geq 0,8$).

Рассмотренная последовательная процедура применялась для выбора наилучшего по критерию среднего числа неопознанных знаков режима работы читающего устройства „РУТА-701“ из нескольких заданных заранее режимов. В случае двух режимов ($s = 2$) и $P^* = 0,9$ выигрыш B был равен 41%.

x
x x

При больших значениях N_0 определение вероятности P_{N_0} и среднего числа испытаний E_{N_0} с использованием рекуррентных формул полученных в пятом разделе связано с трудоемкими вычислениями. В шестой главе предлагается методика определения указанных величин, при использовании которой объем вычислений не зависит от N_0 .

В случае, когда случайные величины X_1, X_2 распределены биномиально ($X_i = 0; 1; i = 1; 2$), методом производящих функций были получены следующие явные выражения для P_{N_0} и E_{N_0} :

$$P_{N_0} = \frac{1}{2d} \sum_{r=1}^d \left[b_r^{N_0} + 2(-1)^{r+1} \left(b_r^{N_0} \sum_{k=d+1}^a e^{d-k} \sin k\varphi_r + e_1^d e_2^{1/2} \frac{1-b_r^{N_0}}{1-b_r} \sin\varphi_r \right) \right] \quad (12)$$

$$E_{N_0} = 2e_2^{1/2} (e_1^d + e_1^{-d}) \sum_{r=1}^d (-1)^{r+1} \sin\varphi_r \cdot \frac{1-b_r^{N_0}}{(1-b_r)^2} \quad (13)$$

где $b_r = P(z=0) + 2e_2^{1/2} \cos\varphi_r$, $\varphi_r = \pi \frac{2r-1}{a+1}$,
 $e_1 = \left(\frac{P(z=-1)}{P(z=1)} \right)^{1/2}$, $e_2 = [P(z=-1)P(z=1)]^{1/2}$,

$$a = 2a_1 + 1, \quad d = a_1 + 1,$$

$$z = \begin{cases} X_2 - X_1 & \text{при } Q_2 > Q_1, \\ X_1 + X_2 & \text{при } Q_2 \leq Q_1, \end{cases}$$

$P(\cdot)$ — вероятность указанного в скобках события.

В случае, когда результаты испытаний X_1, X_2 распределены по закону Пуассона, формулы (12) и (13) являются приближенными формулами для вероятности правильного выбора P_{N_0} и среднего числа испытаний E_{N_0} .

Используя выражения для вероятностей $P(z=j), P(z \geq j)$ и $P(z < j)$ было установлено, что в случае когда значения параметров Q_1 и Q_2 стремятся к нулю, значения вероятности P_{N_0} и среднего числа испытаний E_{N_0} , подсчитанные с использованием формул (12) и (13), стремятся к точным значениям указанных величин.

Исследования, приведенные с использованием цифровой вычислительной машины, показали, что при $Q_1, Q_2 \leq 0,01$ точность предлагаемой аппроксимации достаточна для практических расчетов.

Для облегчения практического использования теоретических результатов, полученных в реферируемой главе, значения величин P_{N_0} и E_{N_0} были табулированы при различных значениях параметров a_1 , N_0 , Q_1 и Q_2 .

x
x x

В седьмой главе исследовались возможности применения предложенных процедур для выбора наилучшего устройства выделения признаков для опознавания речи.

При проведении исследований ставилась задача определить, который из двух устройств выделения признаков речевых сигналов является лучшим по критерию среднего числа случаев ошибочного опознавания. Для проведения экспериментов были выбраны устройства выделения огибающей и мгновенной частоты речевого сигнала разработанные на Кафедре автоматизации производственных процессов Каунасского политехнического института [8, 9].

Устройство для выделения огибающей речевых сигналов было основано на использовании двух линейных четырехполюсников разницы фаз на выходах которых равна 90 градусам. Структурная схема устройства выделения огибающей приводится на рис. 1.

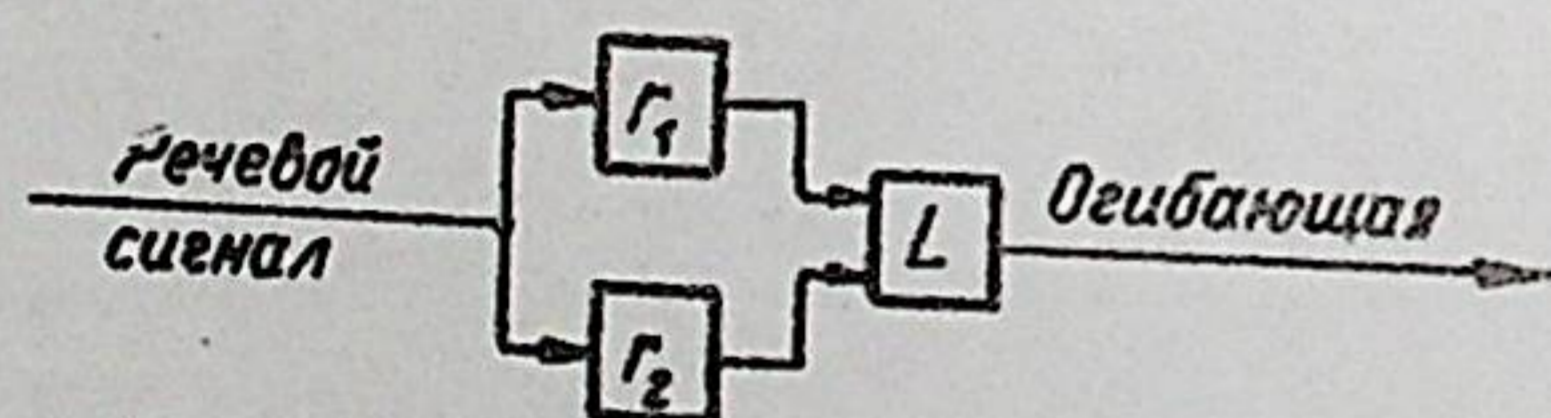


Рис. 1. Структурная схема устройства выделения огибающей речевого сигнала

Здесь:

Γ_1, Γ_2 - линейные четырехполюсники, частотные характеристики которых

$$K_1(\omega)e^{j\psi_1(\omega)}, K_2(\omega)e^{j\psi_2(\omega)}$$

удовлетворяют условиям:

а) $K_1(\omega) = K_2(\omega) = 1,$

б) $\psi_1(\omega) - \psi_2(\omega) = \pi/2$

для $\omega > 0,$

где $\psi_1(\omega), \psi_2(\omega)$ - аргументы коэффициентов передачи первого и второго четырехполюсников, соответственно.

$K_1(\omega), K_2(\omega)$ - модули коэффициентов передачи,

ω - частота,

L - устройство для извлечения квадратного корня из суммы квадратов сигналов, снимаемых на выходах четырехполюсников Γ_1 и Γ_2 .

Для выделения мгновенной частоты применялось устройство, основанное на использовании соотношений, получаемых при смещении по спектру вверх и частотном детектировании речевых сигналов. Смещение спектра вверх осуществлялось известным в технике однополосной радиосвязи фазокомпенсационным методом. Структурная схема устройства приводится на рис. 2.

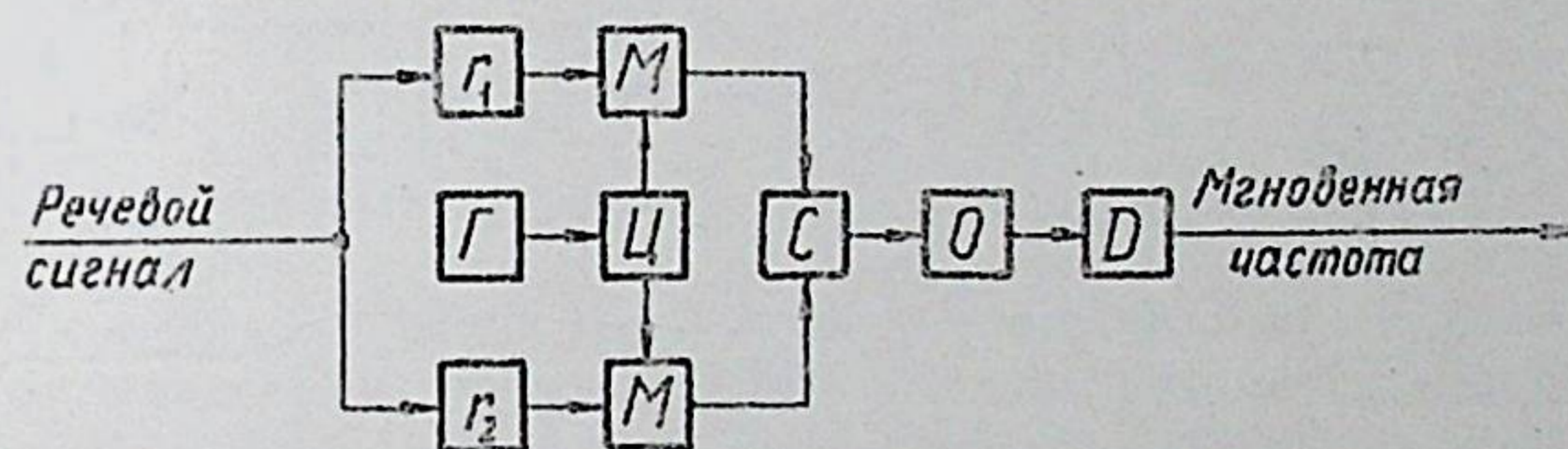


Рис. 2. Структурная схема устройства выделения мгновенной частоты речевого сигнала

Здесь Γ - генератор сигналов стандартной частоты 461,38 кГц, Γ - фазосдвигающая цепочка, на выходах которой снимаются сдвинутые по фазе на $\pi/2$ сигналы стандартной частоты, $М$ - модуляторы, $С$ - сумматор, $О$ - амплитудный ограничитель, $Д$ - частотный детектор.

Для проведения анализа огибающей и мгновенной частоты вышеописанные устройства были соединены с ЭВМ - "Минск 14".

Исследовалась возможность использования описанных устройств для опознавания русских названий 10 арабских цифр: нуль, один, ..., девять. С этой целью в память ЭВМ были введены огибающие и мгновенные частоты 1000 реализаций цифр, произнесенных случайно подобранными дикторами: по 100 реализаций каждой цифры.

Для облегчения экспериментов все реализации были сгруппированы на 20 групп по 50 реализаций в группе. Число испытаний N в этом случае было равно числу групп.

Поскольку частота квантования сигнала была равна 100 гц, а произношение одной цифры занимало около 1 сек., длина каждой реализации составляла около 100 ординат. После ввода в память ЭВМ реализации мгновенной частоты и огибающей известными ме-

годами сокращения описаний, сигналов были описаны векторами с небольшим числом компонент (около 20). Компоненты этих векторов использовались в качестве признаков речевых сигналов.

Для проведения дальнейших исследований были составлены программы, моделирующие работу классификатора опознающего устройства, использующего признаки огибающих и мгновенных частот речевых сигналов. При помощи этих программ моделировался процесс опознавания. Во время моделирования подсчитывались числа X ошибочно опознанных реализаций, из числа 50 реализаций, находящихся в каждой группе. Эти числа X использовались для оценки среднего числа Q ошибочно опознанных слов и для исследования процедур выбора. Значения величины Q приводятся в таблице 5.

Таблица 5

| Число признаков | Мгновенная частота | Огибающая |
|-----------------|--------------------|-----------|
| 15 | 10,4 | 12,5 |
| 20 | 9,8 | 11,35 |

По этим данным были подсчитаны значения вероятности выбора P лучшего устройства выделения признаков (для одновыборочной процедуры). Значения вероятности P так же были оценены экспериментально. Для этого процесс опознавания реализации и выбора лучшего устройства многократно повторялся (100 раз). Часть результатов расчетов и экспериментов для 20 признаков приводится в таблице 6. Здесь: \tilde{P} - экспериментально определенная оценка вероятности P , $P^{(H)}$ и $P^{(B)}$ - верхняя и нижняя доверительные границы, подсчитываемые при уровне значимости 5%.

Таблица 6

| N | P | \tilde{P} | $P^{(H)}$ | $P^{(B)}$ |
|-----|-------|-------------|-----------|-----------|
| 2 | 0,684 | 0,67 | 0,61 | 0,76 |
| 5 | 0,777 | 0,78 | 0,71 | 0,84 |
| 10 | 0,859 | 0,88 | 0,80 | 0,91 |

Для выяснения возможности использования последовательной усеченной процедуры A_y экспериментально были определены оценки \tilde{P}_{N_0} вероятности правильного выбора P_{N_0} и оценки \tilde{E}_{N_0} среднего числа испытаний E_{N_0} . Значения величин P_{N_0} и E_{N_0} так же были подсчитаны теоретически. Часть результатов экспериментов и расчетов - значения величин P_{N_0} , E_{N_0} , \tilde{P}_{N_0} , \tilde{E}_{N_0} в зависи-

мости от параметров N_0 и a_1 приводятся в таблице 7. Здесь: \tilde{G}_{N_0} - оценки среднеквадратических отклонений величины E_{N_0} .

Таблица 7
(20 признаков $Q_1 = 9,8$, $Q_2 = 11,35$)

| a_1 | N_0 | P_{N_0} | $P^{(H)}$ | $P^{(B)}$ | \tilde{P}_{N_0} | E_{N_0} | \tilde{E}_{N_0} | \tilde{G}_{N_0} |
|-------|-------|-----------|-----------|-----------|-------------------|-----------|-------------------|-------------------|
| 10 | 15 | 0,864 | 0,81 | 0,92 | 0,87 | 8,28 | 5,6 | 0,35 |
| | 20 | 0,872 | 0,82 | 0,92 | 0,87 | 8,49 | 5,7 | 0,35 |
| 15 | 15 | 0,897 | 0,85 | 0,94 | 0,91 | 8,94 | 8,52 | 0,40 |
| | 20 | 0,917 | 0,87 | 0,98 | 0,94 | 9,67 | 8,7 | 0,45 |
| 20 | 15 | 0,905 | 0,85 | 0,95 | 0,93 | 10,9 | 11,0 | 0,39 |
| | 20 | 0,931 | 0,89 | 0,97 | 0,97 | 12,2 | 11,8 | 0,47 |

Данные, приведенные в таблицах 6 и 7, показывают хорошее для практических расчетов согласие экспериментальных результатов с теоретическими.

Экспериментальные исследования показали, что теоретические результаты, полученные в настоящей работе могут быть использованы для определения числа реализаций, требуемого для достоверного выбора наилучшего устройства выделения признаков. В случае рассмотренных устройств и заданной вероятности выбора лучшего устройства $P^* = 0,9$ получены следующие результаты: для одновыборочной процедуры при 15 признаках требуются $9 \times 50 = 450$ реализации; а в случае 20 признаков - 750 реализаций.

В случае использования усеченной последовательной процедуры были подсчитаны значения параметра a_1 , при которых гарантируется вероятность $P^* = 0,9$. Так же были подсчитаны средние числа реализаций $50 \cdot E_{N_0}$ и выигрыш B , полученный при использовании рассматриваемой процедуры. Результаты расчетов приведены в таблице 8.

Таблица 8

| Число признаков | N_0 | a_1 | $50 \cdot E_{N_0}$ | $B\%$ |
|-----------------|-------|-------|--------------------|-------|
| 15 | 15 | 10 | 273 | 39 |
| | 20 | 9 | 250 | 44 |
| 20 | 15 | 17 | 487 | 34 |
| | 20 | 13 | 424 | 48 |

Из этой таблицы видно, что требуемое число реализаций, с помощью последовательной процедуры, можно уменьшить на 40%.

В диссертационной работе теоретически и экспериментально исследованы процессы выбора наилучшего (по критерию наименьшей вероятности ошибки опознавания) варианта опознающей системы из заданного множества вариантов. Рассмотрен случай, когда показатели качества вариантов заменяются оценками, вычисленными по выборкам конечных объемов опознаваемых объектов. В случае, когда оценки качества вариантов опознающей системы распределены биномиально или по закону Пуассона, получены следующие научные результаты:

1. Выведены точные и приближенные выражения для числа испытаний опознающей системы, гарантирующего заданную вероятность выбора наилучшего варианта опознающего устройства в случае использования одновыборочной процедуры.

2. Предложена удобная для практического использования последовательная процедура, гарантирующая заданную вероятность правильного выбора, в случае, когда число исследуемых вариантов больше или равно двум.

3. Показано, что вероятность правильного выбора и среднее число испытаний являются возрастающими функциями от коэффициента корреляции между оценками качества вариантов опознающей системы.

4. Установлено, что использование предложений усеченной последовательной процедуры позволяет уменьшить среднее число испытаний до двух раз.

5. Предложена усеченная последовательная процедура выбора наилучшего варианта. Получены выражения для вероятности правильного выбора и среднего числа испытаний.

6. Предложена методика определения значений параметров усеченной последовательной процедуры, гарантирующей заданную вероятность правильного выбора.

7. Установлено, что использование усеченной последовательной процедуры вместо одновыборочной процедуры позволяет уменьшить среднее число испытаний до 40%.

Выражения, полученные в диссертации, приведенные к виду, позволяющему провести практические вычисления на ЭВМ типа БЭСМ-2М. Для облегчения практического использования теоретических результатов основные зависимости табулированы.

Результаты диссертации используются для выбора эффективных режимов работы опознающих устройств в лаборатории читающих автоматов СКБ Вычислительных машин г. Вильнюса и для определения числа реализаций, необходимых для достоверного выбора лучшего устройства выделения признаков речевых сигналов на Ка-

федре автоматизации производственных процессов Каунасского политехнического института.

Исследованные процедуры так же могут быть использованы для решения ряда задач медицинской диагностики, связанных с опознаванием болезней и выбором эффективных способов лечения, поскольку одной из особенностей этих задач является ограниченность числа рассматриваемых опознаваемых объектов опознавания - т.е. людей, больных исследуемыми болезнями. Одной из таких задач, например, является задача выбора комплекса признаков наилучшего по критерию вероятности правильной диагностики сердечных заболеваний.

Основные результаты диссертации опубликованы в следующих работах:

а) Об эффекте зависимости между оценками качества признаков опознающих систем. Вычислительные системы, вып. 28. Новосибирск, 1967.

б) Об одном алгоритме выбора качественной системы признаков. Сб. „Автоматизация ввода письменных знаков в электронно-вычислительные машины“, Вильнюс, 1969.

в) О представительности выборки при решении некоторых задач опознавания образов. В сб. „Техническая кибернетика“. Материалы XIX Респ. научно-технической конф., состоявшейся в 1969 г. в г. Каунасе, Каунас, 1969.

г) Статистическая зависимость при последовательной процедуре выбора качественного процесса опознавания, в сб. „Техническая кибернетика“. Материалы XIX Респ. научно-технической конференции, состоявшейся в 1969 г. в г. Каунасе, Каунас, 1969.

д) Последовательные алгоритмы выбора эффективного режима работы опознающей системы. В сб. „Автоматизация ввода письменных знаков в электронно-вычислительные машины“, Вильнюс, 1969. (Соавтор Р. А. Нашлюнас).

Материалы диссертации докладывались на XVI, XVII и XVIII республиканских научно-технических конференциях, (Каунас, 1966-1969); III Всесоюзной школе-семинаре „Автоматическое распознавание слуховых образов“ (Новосибирск, 1967); научной конференции молодых ученых математиков (Вильнюс, 1968); общесоюзном семинаре „Статистические методы в автоматическом управлении“ (Москва, 1969); II Всесоюзной конференции по автоматизации

зации ввода письменных знаков в ЭВМ (Вильнюс, 1969); IV Сим-
позиуме по экспериментальным задачам (Каунас, 1969).

ЛИТЕРАТУРА

1. Лбов Г. С. „Выбор эффективной системы зависимых признаков“. Вычислительные системы, Новосибирск, Из-во „Наука“ Сиб. отд. 1965, вып. 19.
2. Багдонас А., Жлабис Р., Нашлюнас Р., Яшинскас П. О выборе режима работы системы опознавания, в Сб. „Автоматизация ввода письменных знаков в электронные вычислительные машины“, Изд. „Минтис“, Вильнюс, 1968.
3. Багдонас А., Жлабис Р., Кведеравичюс А., Нашлюкас Р., Шведжидис П., Яшинскас П., „Читающие устройства „Рута-701“, в Сб. „Автоматизация ввода письменных знаков в электронно-вычислительные машины“, изд. РИНТИП, Вильнюс, 1969.
4. Johanson N., On the extension of the connexion between Poisson and x^2 distributions. Biometrika, v. 46, 1959.
5. Abdel-Aty, Approximate formulae for percentage points and the probability integral of the non-central x^2 distribution. Biometrika, v. 41, 1954.
6. Вальд А., Последовательный анализ, ГИФМЛ, М., 1960.
7. Феллер В., Введение в теорию вероятностей и ее применение | „Мир“, 1967.
8. Кемешис П., Читавичус А., Метод формирования огибающей речевых сигналов. Сб. „Автоматика и вычислительная техника“. Труды XV Юбилейной научно-технической конференции, КПИ, Изд. „Минтис“, Вильнюс, 1968.
9. Кемешис П., Читавичус А., Устройство для образования функции мгновенной частоты речевого сигнала. Труды XVI республиканской научно-технической конференции. КПИ, Изд. „Минтис“, Вильнюс, 1969 г.