УДК 621.391

И. Н. ПРЕСНЯКОВ, д-р техн. наук, С. В. ОМЕЛЬЧЕНКО

ПОМЕХОУСТОЙЧИВЫЕ АЛГОРИТМЫ СЕГМЕНТАЦИИ РЕЧИ В СИСТЕМАХ ОБРАБОТКИ

Один из подходов к решению задачи распознавания речи основан на сопоставлении оценок параметров речевых сигналов, найденных по сегментам в виде непрерывной во времени выборки либо последовательности выборок. К таким сегментам можно отнести фонемы, слова, фразы и др. Существующие методы решения задач распознавания речи основаны на изучении структуры речи и её анализе с использованием математического аппарата различного вида и уровня сложности. Исследование структуры речи выполняется также и при решении других прикладных задач, например, в медицинской и криминалистической диагностике, идентификации и верификации диктора и др.

В настоящее время известен ряд математических моделей речеобразования, в основу которых положено разделение речи на вокализованные и невокализованные звуки, а также на паузы между фонемами, слогами, словами и др. Т.е. использование этих моделей, в частности, для решения задач распознавания возможно только после выполнения сегментации речи, что требует нахождения оценок границ между вышеуказанными элементами речи.

Автоматическое разделение речи на сегменты представляет весьма сложную задачу, связанную с существенной нестационарностью речевых сигналов и невозможностью четкой формализации этой задачи в пространстве параметров речевых сигналов таким образом, чтобы сегментация не зависела от конкретного диктора, его эмоционального состояния и других особенностей его голоса. Процедура сегментации также должна обеспечивать инвариантность к случайным изменениям мешающих параметров.

При анализе используются разные критерии для идентификации сегментов, следствием чего является наличие большого числа алгоритмов оценки временных границ сегментов. В связи с этим возникает необходимость в обобщении процедур исследования структуры речи. В настоящее время, однако, неизвестно достаточно эффективное решение данной проблемы.

Целью данной работы является разработка устойчивых к воздействию помех алгоритмов оценивания временных границ сегментов речи.

Рассмотрим математическую постановку задачи сегментации в слитной речи и основные особенности её решения.

Предположим, что на вход системы сегментации поступает временная последовательность отсчетов речевого сигнала $s(l)$, $l = \overline{0, N-1}$ с интервалом дискретизации Δt, задаваемого выражением

$$s(l) = \sum_{k=1}^{p} \alpha_k s(l-k) + u(l), l = \overline{0, N-1},$$

где $u(l)$ - сигнал возбуждения речеобразующего тракта; N- количество временных отсчетов; α_k, $k = \overline{1, p}$ – коэффициенты авторегрессии, описывающей речеобразующий тракт и зависящие от информационного содержания речи диктора.

Априорная информация в виде эталонов сигнала, необходимая для алгоритмов распознавания, задаётся в виде классифицированных обучающих выборок в паузах между словами для каждого из дикторов. Считается, что время появления слова в слитном речевом сигнале априори неизвестно и заданы ограничения на длительность пауз между слогами слов.

Качество $\bar{K}$ алгоритма s будем оценивать величиной дисперсии D(s) оценки временного положения сегментов при отсутствии внешней аддитивной помехи и устойчивостью $k_{уст}(s)$ алгоритма s к воздействию аддитивной помехи

$$\vec{K}(s) = (D(s), k_{уст}(s)).$$

Под показателем устойчивости $k_{уст}(s)$ понимается дисперсия оценки временного положения сегментов при воздействии аддитивной помехи в канале с заданным отношением сигнал/ шум.

Целью рассматриваемой задачи является построение оптимального алгоритма определения по реализациям речи моментов начала и конца сегментов, который обеспечивает максимум целевой функции в классе робастных алгоритмов.

В данной работе в задаче сегментации выделяется два этапа: первый – принятие решения о присутствии речевого сигнала в заданных выборках, второй – оценивание по совокупности выборок временных границ каждого из слов речи.

Первый этап может рассматриваться как бинарная либо многоальтернативная задача, но с последующим принятием двухальтернативного решения.

В алгоритме обнаружения речевой информации, наблюдаемой на фоне гауссовского шума, используется следующее решающее правило:

$$H^0 : \frac{N(\bar{X}^k / \hat{\mu}^1, \hat{R}^1)}{N(\bar{X}^k / \hat{\mu}^0, \hat{R}^0)} < c$$

- принимается гипотеза H^1 о паузе в речевом сигнале;

$$H^1 : \frac{N(\bar{X}^k / \hat{\mu}^1, \hat{R}^1)}{N(\bar{X}^k / \hat{\mu}^0, \hat{R}^0)} \geq c$$

- отвергается гипотеза H^0 о паузе в речевом сигнале.

Здесь с – порог, определяемый в зависимости от выбранного критерия: байесовского ($c = \frac{q(П_{01} - П_{00})}{p(П_{10} - П_{11})}$), максимума апостериорной вероятности ($c = \frac{q}{p}$), максимума правдоподобия ($c = 1$), Неймана-Пирсона (из уравнения $F_{10}(c)=1-\alpha$), минимаксный ($c = \mu_{мм} \frac{(П_{01} - П_{00})}{(П_{10} - П_{11})}$, где параметр $\mu_{мм}$ находится из решения уравнения $\frac{(П_{01} - П_{00})}{(П_{10} - П_{11})} + c(1 - F_{10}(\mu c)) = F_{11}(\mu c)$ при решении указанной задачи).

При гауссовском распределении сигнала с нулевым математическим ожиданием решающее правило сводится к сравнению вычисленных значений с порогом. В случае

$$H^0 : \bar{X}^{(k)tr}[(\hat{R}^1)^{-1} - (\hat{R}^0)^{-1}]\bar{X}^{(k)} > \hat{\Delta}^0$$

– принимается гипотеза о паузе в речевом сигнале;

$$H^1 : \bar{X}^{(k)tr}[(\hat{R}^1)^{-1} - (\hat{R}^0)^{-1}]\bar{X}^{(k)} \leq \hat{\Delta}^0$$

– отвергается гипотеза о паузе в речевом сигнале k-й выборки.

Здесь $\hat{R}^0$, $\hat{R}^1$ – полученные по обучающей выборке оценки корреляционных матриц шумов в паузах речи и речевого сигнала.

Для стационарных случайных процессов выполняется условие δ-коррелированности спектральных коэффициентов $c_i^{(k)}$ в базисе комплексно-экспоненциальных функций. При этом гипотеза о паузе в речевом сигнале принимается, если

$$H^0 : \sum_{i=0}^{N} \left| c_i^{(k)} \right|^2 [(S_i^{(1)})^{-1} - (S_i^{(0)})^{-1}] > \hat{\Delta}_s{}^0.$$

Гипотеза о паузе в речевом сигнале k-ой выборки отвергается, если

$$H^1 : \sum_{i=0}^{N} \left| c_i^{(k)} \right|^2 [(S_i^{(1)})^{-1} - (S_i^{(0)})^{-1}] \le \hat{\Delta}^0 .$$

Здесь $S_i^{(1)}$ – полученные по обучающей выборке оценки корреляционной функции спектральных коэффициентов.

В случае, если форма энергетического спектра шума и речевого сигнала идентичны, т.е. $S_m^{(1)} = uS_m^{(0)}$, то $[(S_i^{(1)})^{-1} - (S_i^{(0)})^{-1}] = -a / S_m^{(0)} = H(m)H(m)^*$, где $a = (u-1)/u$ - постоянный коэффициент, $H(m)$- передаточная функция выбеливающего фильтра. В случае, если форма энергетического спектра шума и речевого сигнала различны и $(S_i^{(1)}) >> (S_i^{(0)})$, то выполняется приближение $[(S_i^{(1)})^{-1} - (S_i^{(0)})^{-1}] \approx -(S_i^{(0)})^{-1} = = H(m)H(m)$.

Если речевой сигнал пропустить через выбеливающий фильтр, алгоритм сегментации сводится к алгоритму временной обработки некоррелированых отсчетов в выборках. В последующем изложении считается, что выборки речевого сигнала, используемые в алгоритмах обработки, некоррелированны. Для устранения корреляции используется декоррелирующие преобразования, например, разложение по методу Холецкого [3] обратной корреляционной матрицы речевого сигнала на произведения нижнетреугольной и верхнетреугольной матриц.

Другим способом декорреляции речевого сигнала является его разложение по базису Карунена-Лоэва. При этом решающее правило получаем в виде

$$H^1 : \sum_{j=1}^{N} c_j^{(k)2} / \hat{\lambda}_j^{(k)} \le \hat{\Delta}_c^0, \qquad H^0 : \sum_{j=1}^{N} c_j^{(k)2} / \hat{\lambda}_j^{(k)} > \hat{\Delta}_c^0,$$

где $\hat{\lambda}_j = (\hat{\sigma}_{jc}^0)^2$ – собственные числа выборочной корреляционной матрицы сигнала; $c_j^{(k)}$ – представление к-й реализации сигнала $\hat{X}$ в базисе Карунена-Лоэва с размерностью N. Собственные числа и собственные векторы вычисляют в виде $\Phi^*[(\hat{R}^1)^{-1} - (\hat{R}^0)^{-1}]\Phi = \Lambda$, а в случае $\hat{R}^1 = s\hat{R}^0$, s>>1 вычисление упрощается к виду $\Phi^*(\hat{R}^0)^{-1}\Phi = \Lambda$.

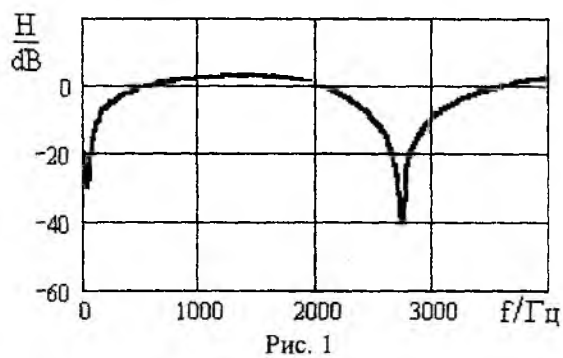

Рис. 1

Полагая, что в пределах выборки речевой сигнал стационарен в широком смысле, алгоритм выбеливания речевого сигнала в частотной области имеет вид

$$\hat{s}(t) = \mathrm{Re}\left(\frac{1}{\sqrt{2N}} \sum_{m=0}^{2N-1} C(m)H(m)\exp(i(2\pi t / N)m) \right),$$

$$C(\omega) = \frac{1}{\sqrt{2N}} \sum_{\tau=0}^{2N-1} y_\tau^j \exp(-i(2\pi\tau\omega / N)),$$

где $y_i^j = \begin{cases} x_i^j, & i = 0,1,...,(N-1) \\ 0, & i = N,(N+1),...,(2N-1) \end{cases}$ – входные отсчеты; $H(m) = A / \sum_{l \in Z} W(l)(S(m+l))^q$ - амплитудно-частотная характеристика выбеливающего фильтра;

$S(m)=\left|\frac{1}{\sqrt{N}}\sum_{\tau\in Z}K(\tau)\exp(-i(2\pi m\tau/N))\right|$ – оценка энергетического спектра;

$K(\tau)=\frac{1}{(T+1-\tau)L}\sum_{j=1}^{L}\sum_{i=0}^{T-\tau}x_{i+\tau}^{(j)}\,x_i^{(j)}$ – оценка корреляционной функции речевого сигнала.

На рис. 1 приведена амплитудно-частотная характеристика фильтра $H(m)$ для случая, когда $W(0)=1, W(1)=1$, $W(j)=0$, где $j\in Z; j\neq 0; j\neq 1$.

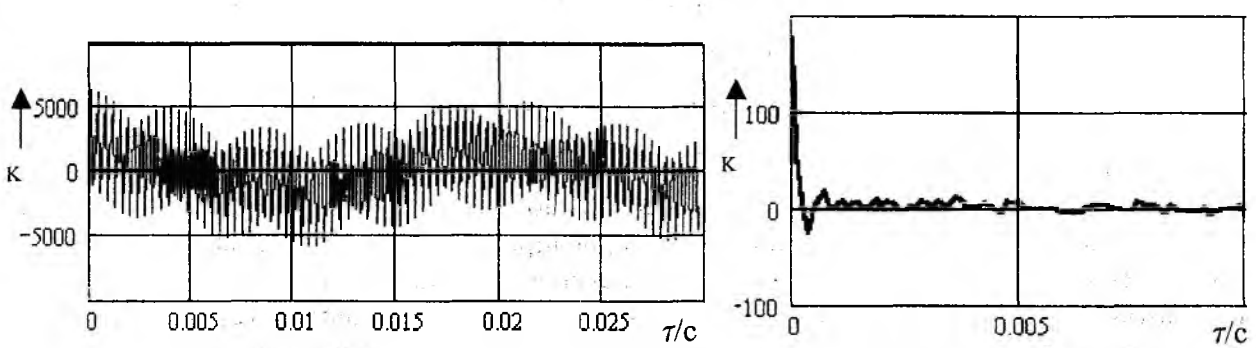

Рис. 2

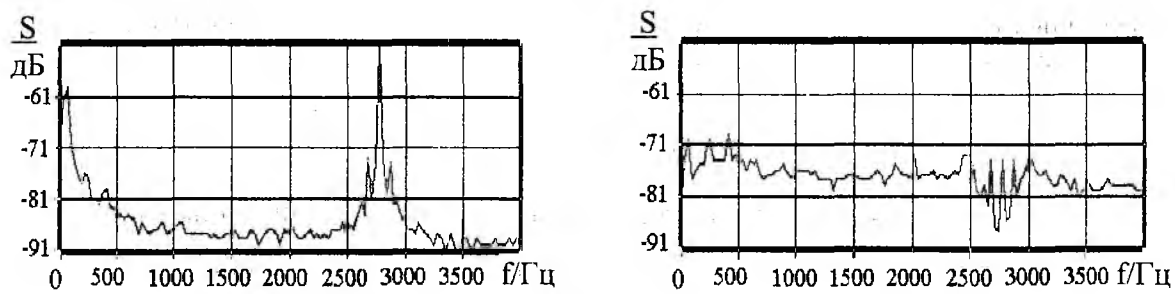

Рис. 3

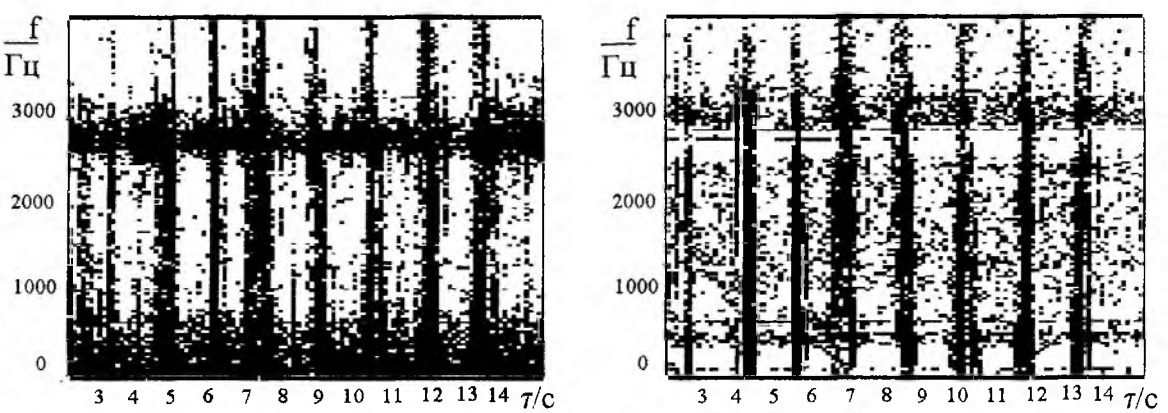

Рис. 4

На рис. 2а приведена корреляционная функция речевого сигнала в паузе до фильтрации, на рис. 2б – корреляционная функция речевого сигнала в паузе после выбеливания речевого сигнала.

Экспериментальные исследования речевых сигналов показали, что одномерный в пространстве параметров частот энергетический спектр сигнала в паузе, полученный усреднением 20 выборок по 256 отсчетам (см. рис. 3а), существенно отличается от равномерного, т.е. шум не является белым. На рис. 3 б приведен энергетический спектр речевого сигнала в

паузе, полученный в результате выбеливания сигнала фильтром. На рис. 4а, б приведены амплитудно-временно-частотные характеристики наблюдаемого речевого сигнала соответственно до и после выбеливания речевого сигнала фильтром .

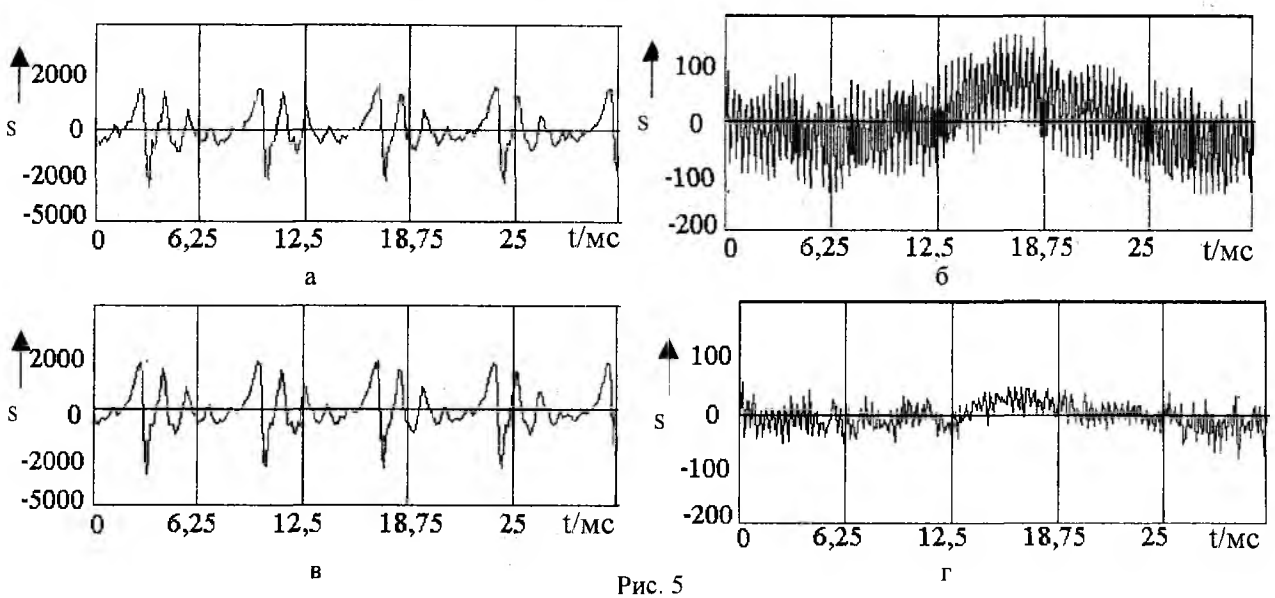

Рис. 5

Как показали исследования, использование такого фильтра позволяет существенно повысить соотношение сигнал-шум, что обусловливает более высокое качество обнаружения речи для ряда рассматриваемых ниже алгоритмов распознавания. На рис. 5 а,б приведены соответственно выборки вокализованного речевого сигнала и шума в паузе в случае отсутствия выбеливания. На рис. 5 в,г показаны выборки вокализованного речевого сигнала и шума в паузе после выполнения процедуры выбеливания.

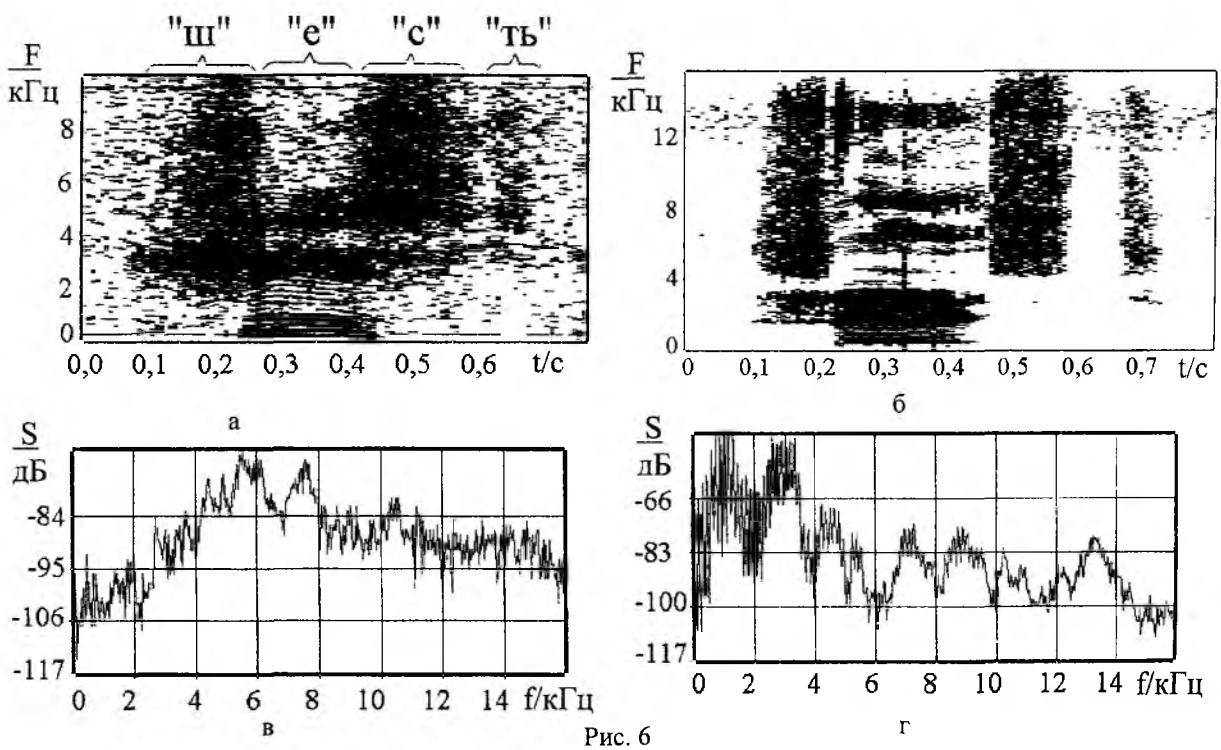

Рис. 6

Повышение качества сегментации речи может быть достигнуто фильтрацией речевого сигнала фильтром с частотной характеристикой $H(m)$, удовлетворяющей условию

$$\frac{1}{S_m^{(1)}} - \frac{1}{S_m^{(0)}} = H(m)H(m)^*, \text{ откуда } \frac{1}{S_m^{(0)}} \frac{S_m^{(0)} - S_m^{(1)}}{S_m^{(1)}} = H(m)H(m)^*.$$

Такая коррекция частотной характеристики наиболее эффективна в случае слабых сигналов, т.е. при сопоставимых значениях $S_m^{(0)}$ и $S_m^{(1)}$.

Таким образом, при построении алгоритмов принятия решений необходимо учитывать не только необходимость введения обеляющего фильтра, но и формантную структуру энергетического спектра речевого сигнала. На рис. 6 а, б приведены амплитудно-временно-частотные характеристики слова «шесть» для двух дикторов. Они подтверждают характерные особенности вокализованных звуков, заключающихся не только в линейчатости спектра, вызванной периодичностью речевого сигнала, но и в преобладании энергии спектра в низкочастотной полосе до 3,4-5 кГц (см. рис. 6 г – энергетический спектр для вокализованного звука «е»), а для невокализованных звуков – преобладание энергии спектра в полосе от 1,5-4 кГц до 6-20 кГц (см. рис. 6 в – энергетический спектр для невокализованного звука «с»).

Для нахождения оптимальной оценки величины порога и длительности выборки, по которой принимается решение, рационально использовать адаптивные алгоритмы.

В результате применения выбеливающего фильтра, который может быть реализован как нерекурсивный фильтр либо как фильтр в частотной области, алгоритмы обнаружения могут быть упрощены за счет декорреляции временных отсчетов речевого сигнала.

При обеспечении некоррелированности признаков и равенства дисперсий в координатном представлении в алгоритме обнаружения речевого сигнала по энергетическим признакам выносится решение о наличии речевой информации в к-й выборке, если выполняется неравенство

$$H_1 : (l(k)) > \Lambda, \tag{1}$$

где $l(k) = \sum_{i=1}^{N} \left| S_i^k \right|^2$, а S_i^k -1-й отсчет k-й выборки речевого сигнала.

В противном случае выносится решение о наличии паузы.

Порог Λ в общем случае вычисляется, как $\Lambda = \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln\left[\left(\frac{\sigma_1^2}{\sigma_0^2}\right)^n c\right] > 0$, $\sigma_1 > \sigma_0$. (1a)

Для критерия Неймана-Пирсона при заданном α порог (1a) преобразуется к виду $\Lambda = \sigma_0^2 \chi_\alpha^2$, где χ_α^2 – выраженное в процентах отклонение случайной величины, распределенной по закону $\chi 2$ с n степенями свободы.

Рассмотрим особенности сегментации речи в пространстве оценок ковариационных матриц с распределением Уишарта

$$W(\hat{K} / R) = \frac{\det\left|\hat{K}\right|^{\frac{\tau-p-1}{2}} \exp\left(-\frac{sp\left(\hat{K}R^{-1}\right)}{2}\right)}{\det\left|R\right|^{\frac{\tau}{2}} \gamma(\rho, \tau)}.$$

При априорном знании ковариационных матриц R_1, R_2 процедура принятия решения состоит в сравнении логарифма отношения правдоподобия E_n с порогом $k = \ln(c)$ априорно выбранного критерия качества.

Гипотеза о наличии сигнала с ковариационной матрицей R_1 принимается, если выполняется неравенство

$$E^r = \ln W(\hat{K}^r / R_2) - \ln W(\hat{K}^r / R_2) = 0{,}5sp\left(\hat{K}^r\left[R_1^{-1} - R_2^{-1}\right]\right) \geq ln\, c - \frac{\tau}{2}\ln(\det R_1 / \det R_2).$$

Поэтому алгоритм сегментации по множеству оценок ковариационных матриц сводится к виду

$$\sum_{l=0}^{n-1}\sum_{i=0}^{n-1} h(j,i)\hat{K}^r(i,j) \geq \Lambda_1, \text{ где } h(j,i) = R_1^{-1}(j,i) - R_2^{-1}(j,i).$$

В пространстве оценок энергетического спектра $S^k(i)$ стационарного случайного процесса алгоритм сегментации имеет вид

$$\sum_{i=0}^{N-1} S^k(i)H(i) \geq \Lambda.$$

Из результатов экспериментальных исследований (рис.6в,г)следует, что средние частоты энергетических спектров смещены в область высоких частот для невокализованных звуков и в область низких частот для вокализованных звуков в сравнении со средней частотой, характерной для белых шумов.

Принятие решения возможно по степени близости оценки энергетического спектра к эталону. Мера близости выбирается как расстояние в гильбертовом пространстве между оценкой и эталоном энергетического спектра для данного типа речевого сигнала

$$\sum_{i=0}^{N-1}\left(S^k(i) - H^j(i)\right)^2 \leq \sum_{i=0}^{N-1}\left(S^k(i) - H^l(i)\right)^2,$$

где $S^k(i)$ – оценки энергетического спектра, $H^j(i)$ – эталоны, вычисленные в результате усреднения энергетического спектра для заданного j типа речевого сигнала, например, для вокализованного и невокализованного речевого сигнала, а также шума в паузе.

В случае распознавания по форме энергетического спектра квадратичное решающее правило преобразуется в линейное вида

$$\sum_{i=0}^{N-1} S^k_{norm}(i)H(i) \geq \Lambda,$$

где $S^k_{norm}(i) = \dfrac{S^k(i)}{\sum_{i=0}^{N-1}\left|S^k(i)\right|}$ – нормированная оценка энергетического спектра речевого сигнала, $H(i) = H^j(i) - H^l(i)$ - весовой коэффициент, $\Lambda = \sum_{i=0}^{N-1}\left(H^j(i)\right)^2 - \sum_{i=0}^{N-1}\left(H^l(i)\right)^2$ – оценка порога решающего правила.

В алгоритмах обнаружения речевого сигнала выносится решение о наличии вокализованного фрагмента речевого сигнала в к-й выборке, если выполняется неравенство

$$H_2 : (l_B(k)) > \Lambda_1$$

и решение о наличии невокализованного фрагмента речевого сигнала, если

$$H_1 : (l_n(k)) > \Lambda_2,$$

где $l_B(k) = \sum_{i=0}^{N-1} H_B(i)\left|S^k_{norm}(i)\right|$, $l_n(k) = \sum_{i=0}^{N-1} H_n(i)\left|S^k_{norm}(i)\right|$, при этом $S^k(i)$ -i-я составляющая

оценки энергетического спектра для k-й выборки, $H_B(i), H_n(i)$ – весовые коэффициенты для вокализованных и невокализованных выборок речевого сигнала.

Из алгоритма обнаружения по оценке формы энергетического спектра при линейной аппроксимации весовых коэффициентов $H_B(i) = N - i$, $H_n(i) = i$ получим алгоритм обнаружения речевой информации по оценкам средней частоты спектра.

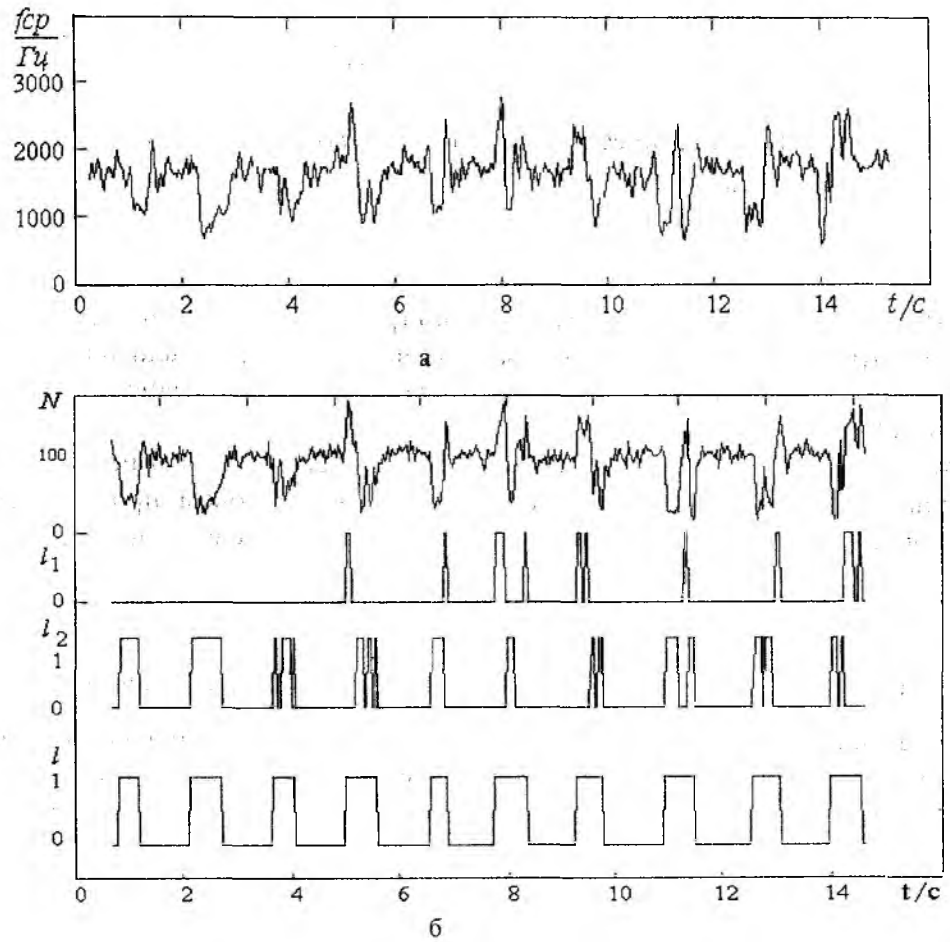

Рис. 7

Алгоритм обнаружения по оценкам средней частоты спектра выносит решение о наличии вокализованного фрагмента речевого сигнала в к-й выборке, если выполняется неравенство

$$H_2 : (l(k)) < \Lambda_1$$

и решение о наличии невокализованного фрагмента, если

$$H_1 : (l(k)) > \Lambda_2, \qquad (2)$$

при этом оценка средней частоты энергетического спектра вычисляется как

$$l(k) = \frac{1}{b+a+1} \sum_{v=k-a}^{k+b} \frac{\sum_{i=0}^{N-1} i \left| S^v(i) \right|}{\sum_{i=0}^{N-1} \left| S^v(i) \right|},$$

где $S^k(i)$ -i-я составляющая оценки энергетического спектра для k-й выборки.

На рис. 7а построены траектории оценки средней частоты энергетического спектра для последовательности из 460 выборок речевого сигнала по 256 отсчетам. При этом средняя

частота находилась независимо для каждой выборки. При построении траектории производилось сглаживание оценок двух соседних выборок.

В алгоритмах обнаружения по признакам взвешенной разности энергий выносится решения о наличии вокализованного фрагмента речи в к-й выборке, если выполняется неравенство

$$H_2 : (l(k)) < \Lambda_1$$

и невокализованного фрагмента речи, если

$$H_1 : (l(k)) > \Lambda_2 . \qquad (3)$$

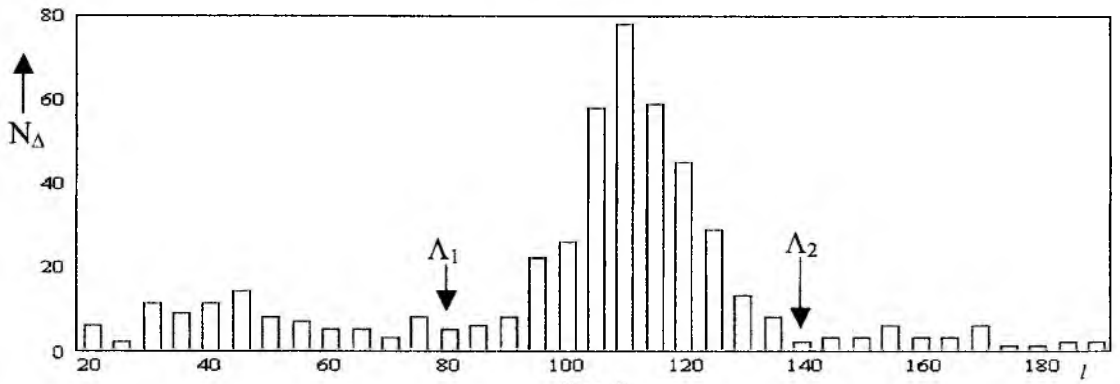

Рис. 8

При этом для оценки разности взвешенных энергий речевого сигнала в выделенных полосах частот аппроксимация весовых коэффициентов задается в виде

$$H_B(i) = \begin{cases} a, \text{ где } 0 \le i < N1, \\ -b, \text{ где } N1 \le i < N-1, \end{cases} \text{ а } H_n(i) = \begin{cases} -a, \text{ где } 0 \le i < N1, \\ b, \text{ где } N1 \le i < N-1, \end{cases}$$

где $a = N / N1$, $b = N /((N - N1)$ вычисляют по формуле

$$l(k) = -\frac{N}{N1} \sum_{i=0}^{N1-1} \left| S^k_{norm}(i) \right| + \frac{N}{N - N1} \sum_{i=N1}^{N-1} \left| S^k_{norm}(i) \right| .$$

Оценку средней частоты энергетического спектра речевого сигнала можно получить по числу пересечений речевым сигналом нулевого уровня в пределах выборок.

В алгоритмах обнаружения речевого сигнала в выборках по признакам нуль-пересечения выносится решение о наличии вокализованного фрагмента речи в к-й выборке, если выполняется неравенство

$$H_2 : (l(k)) < \Lambda_1$$

и невокализованной, если

$$H_1 : (l(k)) > \Lambda_2 . \qquad (4)$$

Вычисление пороговых уровней Λ_1, Λ_2 производится по результатам определения локальных минимумов близлежащих справа и слева от глобального максимума гистограммы распределения решающей статистики (см. рис. 8).

На рис. 9 приведена зависимость оценки числа пересечений нулевого уровня в пределах выборки от номера выборки k, а также вычисленные оценки уровней порога Λ_1, Λ_2, необходимые для принятия решений о наличии вокализованных и невокализованных участков речевых сигналов в выборках.

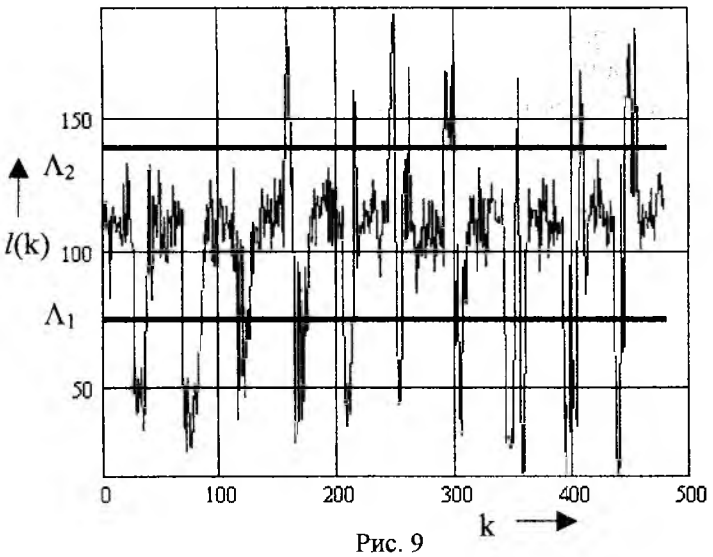

Рис. 9

В алгоритме обнаружения речевого сигнала в выборках по признакам периодической коррелированности выносится решение о его наличии в к-й выборке, если

$$l(k,\tau) < \Lambda_1, \qquad (5)$$

$$l(k,\tau) = \max_{T \in T_{ljg}} \sum_{i=1}^{N} S_i^k S_{i-\tau}^k \cos(\frac{2\pi}{T} l),$$

где S_i^k – i-й отсчет k-й выборки после нормирования по энергии.

Временной сдвиг τ может выбираться экспериментально с целью нахождения максимума величин $l(k,\tau)$.

Для сегментации возможно использование формантных (модифицированных авторегрессионных) оценок, которые вычисляются в соответствии с выражением

$$\vec{f}_v = \frac{F_\partial}{N} \arg loc \max \left\{ \left| 1 + \sum_{n=1}^{p-1} \left(\hat{a}[n] \exp(-j2\pi nk) + \alpha \exp(-j2\pi pk) \right) \right|^{-1}, k = \overline{0, M} \right\},$$

где α-коэффициент, близкий к единице (например $\alpha = 0{,}99$, при $\alpha = 1$ алгоритм становится неустойчивым);

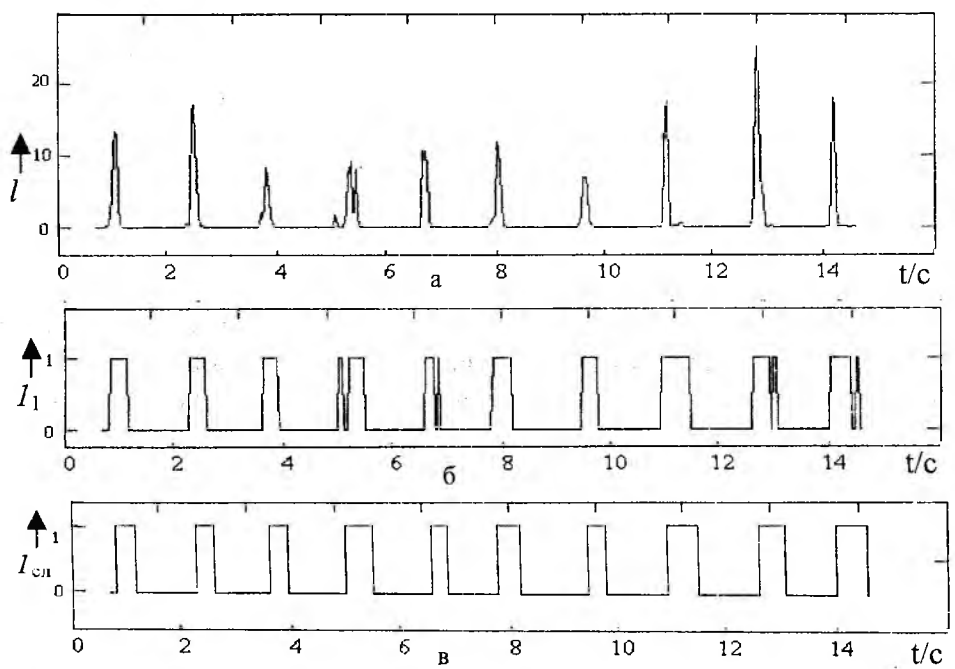

Рис. 10

$\vec{f} = \arg loc \max(\vec{x})$ – векторная функция, задающая соответствие элементам входной последовательности $x_1, x_2, \ldots, x_N$ элементам выходной последовательности упорядоченное множество номеров локальных максимумов $\{ f_i, i = \overline{0, L} \}$.

Решение о начале нового сегмента фонем в очередной выборке принимается по результату сравнения с порогом значений $R_n^{фон}$, вычисленных по формуле (6)

$$R_n^{фон} > \Lambda,$$

где R_n – функционалы, построенные на основе метрик в пространстве L1, L2

$$R_n^{фон} = \sum_{i=1}^{L(n)} \min_{j\in[-J,J]} \alpha_{i,j}^{l} \left| \hat{f}_i(n) - \hat{f}_{i+j}(n+1) \right|^q, \qquad (6)$$

где $\hat{f}_i(n)$ – оценки частот i-й форманты n-го сегмента; $\alpha_{i,j}^{l}$ – весовые коэффициенты, $i = \overline{-J,J}$; $j = \overline{-J,J}$; q принимает значения 1 или 2 в зависимости от вида критерия близости.

На основе первичной сегментации слов по формантным признакам выносится решение о наличии речевой информации в n-м сегменте в случае, если

$$H_1 : R_n^{слов} < \Lambda,$$

$$R_n^{слов} = \sum_{i=1}^{L(n)} \min_{j\in[-J,J]} \alpha_{i,j}^{l} \left| \hat{f}_i(n) - \hat{f}_{i+j}^{П} \right|^q, \qquad (7)$$

где $\hat{f}_i(n)$ – оценки частот i-й форманты n-го сегмента; $\hat{f}_{i+j}^{П}$ – эталонные оценки частот i-й форманты, полученные усреднением оценок для нескольких сегментов, соответствующих паузе речи; $\alpha_{i,j}^{l}$, $i = \overline{-J,J}$; $j = \overline{-J,J}$ – весовые коэффициенты.

На втором этапе обработки вычисляются границы слов по принципу временной компактности с исключением пауз не более заданной длительности (см. рис. 10 – энергетическая оценка границ слова, рис. 7б – оценка границ слова, полученная из объединения информации вокализованных и невокализованных участков слова). На рис. 10а показана зависимость энергетической статистики (1) от номера выборки, на рис.10б – решение бинарной задачи обнаружения, а на рис. 10в – результат оценивания границ слов после исключения пауз между слогами. На верхней эпюре рис. 7б показано изменение числа N нуль-пересечений от номера выборки, на двух следующих – результат принятия решения бинарной задачи обнаружения l_1 вокализованных и l_2 невокализованных сегментов речи, а на нижней эпюре – результат оценивания границ слов после объединения информации вокализованных и невокализованных участков слова и исключения пауз между слогами.

Рассмотрим результаты экспериментального исследования алгоритмов сегментации речи. Исследования описанных выше методов сегментации выполнены по выборкам реальных речевых сигналов для разных дикторов. Оценивание производилось в соответствии с алгоритмами (1)- (7).

С целью звукового контроля качества сегментации речи с помощью экспертов проведены экспериментальные исследования. По отсчетам звукового сигнала, следующих в результате дискретизации с частотой 8 кГц, производились оценки временных границ начала и конца каждого из 10 слов речи. Результаты оценивания границ слов (слогов) в соответствии с алгоритмами 1,4-7 использовались для выделения фрагментов активной речи с последующей её записью на диск в формате «PCMWAVEFORMAT» с расширением «wav». Выделение фрагментов активной речи из звукового сигнала s(t) происходит в соответствии с алгоритмом

$$s_2(t) = s(t) \sum_{p=1}^{P} h_p(t),$$

где $h_p(t) = \begin{cases} 1, & t \in [t_{нр}, t_{кр}] \\ 0, & t \notin [t_{нр}, t_{кр}] \end{cases}$, p – текущий номер слова, $t_{нр}, t_{кр}$ – оценки моментов времени, соответствующие началу и концу слова (слога).

Звуковой контроль воспроизводимой речи по исследуемым алгоритмам (1,4-6) не позволил выявить дефекты речи, связанные с пропаданием отдельных фонем или др. составных частей слов.

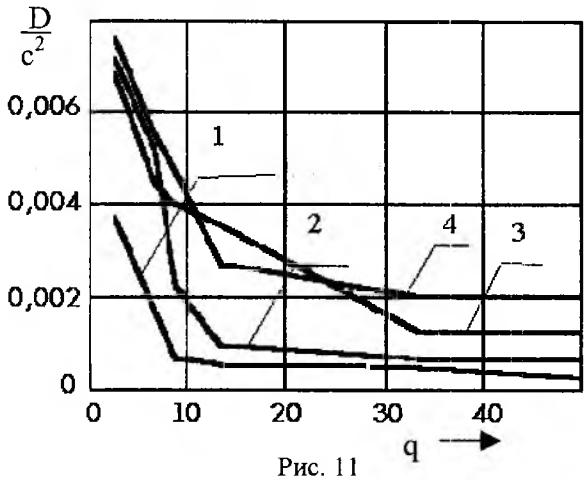

Рис. 11

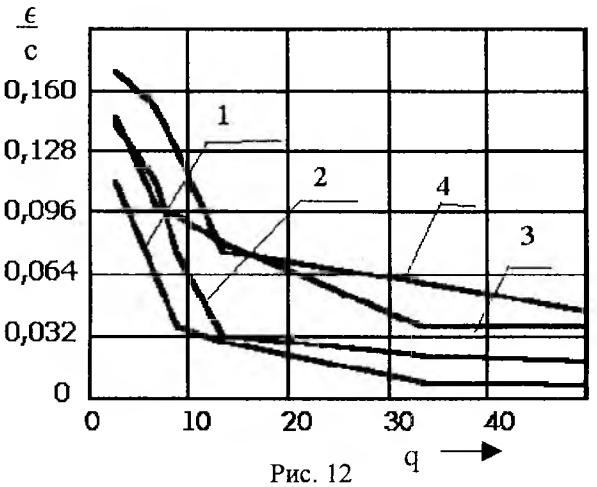

Рис. 12

С целью изучения влияния выбора частоты дискретизации речевого сигнала, уровня аддитивной помехи на результирующие показатели качества проектируемого устройства сегментации речевого сигнала экспериментально исследован ряд зависимостей. Исследована дисперсия оценивания временных границ сегментов речи для алгоритмов (1,4-6) (кривые дисперсии 1-4 рис. 11), а также экспериментально получены кривые смещений оценок временного положения начала и конца слова (1-4 рис. 12) от отношения сигнал/шум q при дли-

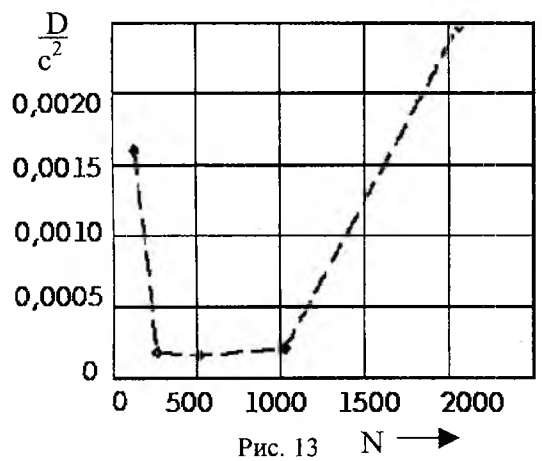

Рис. 13

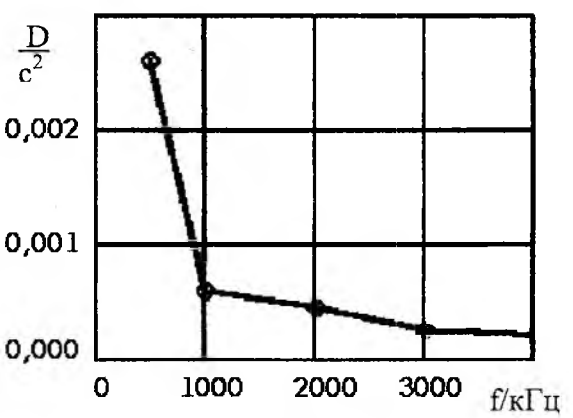

Рис. 14

тельности выборки 256 отсчетов и частоте дискретизации речевого сигнала 8 кГц. Зависимость дисперсии оценивания для алгоритма сегментации слов по признакам нуль- пересечения от длины выборки N приведена на рис. 13. В результате исследований получено, что среднеквадратическое отклонение оценок временных границ слов алгоритма сегментации слов по признакам нуль-пересечения с выбеливанием будет не более 360 отсчетов при длительности выборки от 256 до 1024 отсчетов и частоте дискретизации речевого сигнала 8 кГц.

Экспериментально получена зависимость дисперсии оценивания от верхней частоты спектра сигнала для алгоритма (1) с выбеливанием (см. рис. 14).

Таблица 1

Алгоритмы сегментации слов	D, $с^2$	D,$с^2$ при q=13
По энергетическим признакам с выбеливанием	0,00025	0,00053
По энергетическим признакам без выбеливания	0,0024	0,0036
По средний частоте энергетического спектра с выбеливанием	0,00099	0,0048
По разности энергий в НЧ и ВЧ полосах речевого сигнала с выбеливанием	0,00115	0,0054
По признакам нуль пересечения с выбеливанием	0,00123	0,00354
По признакам периодической коррелированности для τ=1	0,00063	0,00094
Формант (модифицированный) для порядка модели 12	0,00206	0,0027

В табл. 1 приведены результаты исследования 7 вариантов устройств сегментации слов, отличающихся типом алгоритма оценивания начала и конца слова. Из таблицы видно, что тип устройства в смысле критерия максимума дисперсии оценивания временного положения слова зависит от требований устойчивости. Если задать допустимое значение показателя устойчивости $К_{уст}(s)$, соответствующего отношению сигнал шум q=13, то наилучшим по показателю дисперсии оценивания временного положения D будет алгоритм сегментации слов по энергетическим признакам с выбеливанием.

Заключение. В настоящей работе разработаны устойчивые алгоритмы оценивания временных границ слов на основе: признаков нуль-пересечений, энергетических, компонентных статистик в рамках модели ПКСП, формантных признаков. На основе линейной модели речевых сигналов синтезированы и исследованы алгоритмы слоговой сегментации, сегментации на вокализованные и невокализованные фрагменты речи, а также фонемы на основе методов нуль-пересечений и формантных признаков. По результатам экспериментальных исследований, выполненных на реальных сигналах, показана возможность практического использования рассмотренных алгоритмов сегментации речевых сигналов. Найдены характеристики приведенных алгоритмов. Приведены результаты экспериментальных исследований качества алгоритмов сегментации для заданных выборок речевых сигналов при воздействии дополнительного аддитивного гауссова белого шума. Исследованы различные возможные пути решения сформулированной задачи.

Список литературы: 1. *Дж. Д. Маркел, А. Х. Грей.* Линейное предсказание речи. М.: Связь, 1980. 308 с. 2. *Рабинер Л. Р., Шафер Р. В.* Цифровая обработка речевых сигналов / Под ред. М. В. Назарова и Ю. Н. Прохорова. М.: Радио и связь, 1981. 496 с. 3. *Марпл. – Мл. С. Л.* Цифровой спектральный анализ и его приложения. М.: Мир, 1990. 584 с.

Харьковский национальный
университет радиоэлектроники

Поступила в редколлегию 15.12.2002