УДК 621.391

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

РАСПОЗНАВАНИЕ РЕЧЕВОГО СИГНАЛА НА ФОНЕ КОРРЕЛИРОВАННОЙ ПОМЕХИ

Линейное предсказание на основе решетчатых (лестничных) фильтров является одним из наиболее эффективных методов анализа речевого сигнала. Этот метод становится доминирующим при оценках функций площадей сечений голосового тракта. Важность метода обусловлена высокой точностью получаемых оценок и относительной простотой вычислений. Представление коррелированных случайных сигналов моделями в виде решетчатых фильтров находит широкое практическое применение [4–8]. Так, решетчатые фильтры широко применяются при спектральном оценивании случайных сигналов, сегментации речи [4]. Особое практическое значение модели линейного предсказания имеют при создании эффективных алгоритмов распознавания речи на основе результата выбеливания решетчатых фильтров, коэффициентов отражения и логарифмов отношения площадей сечений голосового тракта.

Целью статьи является построение алгоритмов распознавания речи в условиях действия узкополосных помех.

Рассмотрим математическую постановку задачи распознавания слов речи.

Полагается, что на вход системы распознавания поступает временная последовательность отсчетов речевого сигнала x_n, $n = \overline{1, N}$, взятых с интервалом дискретизации Δt.

Для создания алгоритмов распознавания важны априорные сведения о вводимых словах. Эталоны слов для каждого из дикторов заданы в виде классифицированных обучающих выборок.

Считается, что время предъявления речевых единиц в речевом сигнале априори не известно. Положим, что априорные вероятности предъявления для всех слов речи одинаковы.

Необходимо синтезировать алгоритм, который по предъявленной реализации речи выносит решения о конкретном слове и обеспечивает максимум средней вероятности правильного распознавания слов, при этом средняя вероятность правильного распознавания слов речи не менее заданной при воздействии аддитивной помехи в канале связи с заданным отношением сигнал-шум q.

Решение задачи распознавания речи и ряда других задач во многом связано с успешным проведением сегментации речи на речевые единицы. Подобная сегментация на этапе распознавания речевых единиц позволяет исключить избыточные процедуры принятия решений по сигналам, не несущим речевую информацию либо не являющимися целостными речевыми единицами. Задача сегментации состоит в членении речи на структурные единицы и оценивании их временных границ. Некоторые алгоритмы сегментации описаны и исследованы в [1–4].

Рассмотрим предварительную обработку речевого сигнала фильтром. В лестничном методе коэффициенты предсказания оцениваются непосредственно по речевому сигналу без промежуточного вычисления автокорреляционной функции.

Коэффициенты отражения (коэффициенты частной корреляции) могут вычисляться через ошибки предсказания $d^m{}_n$, $b^m{}_n$ в соответствии с выражением

$$k_m = \frac{\sum_{n=1}^{N} (d^{m-1}{}_n \cdot b^{m-1}{}_{n-1})}{(\sum_{n=1}^{N} (d^{m-1}{}_n)^2 \cdot \sum_{n=1}^{N} (b^{m-1}{}_{n-1})^2)^{1/2}}, \qquad (1)$$

где $m = 1, \ldots, p$.

Коэффициенты отражения согласно методу максимальной энтропии Бурга

$$k_m = \frac{2 \cdot \sum_{n=1}^{N} (d^{m-1}{}_n \cdot b^{m-1}{}_{n-1})}{\sum_{n=1}^{N} (d^{m-1}{}_n)^2 + \sum_{n=1}^{N} (b^{m-1}{}_{n-1})^2}. \quad (2)$$

Полагается, что начальные ошибки прямого $d^0{}_n$ и обратного предсказания $b^0{}_n$ имеют вид

$$d^0{}_n = x_n, \quad b^0{}_n = x_n, \quad (3)$$

где $n = 1,\ldots, N$.

Ошибки прямого $d^m{}_n$ и обратного предсказания $b^m{}_n$ могут быть представлены в виде

$$d^m{}_n = d^{m-1}{}_n - k_m \cdot b^{m-1}{}_{n-1},$$

$$b^m{}_n = b^{m-1}{}_{n-1} - k_m \cdot d^{m-1}{}_n. \quad (4)$$

Результат работы выбеливающего фильтра имеет вид

$$y_n = d^p{}_n. \quad (5)$$

В наших экспериментальных исследования коэффициенты частной корреляции и коэффициенты отражения согласно методу максимальной энтропии Бурга дали одинаковые результаты оценивания коэффициентов отражения.

На рис.1 показано устройство вычисления коэффициентов отражения.

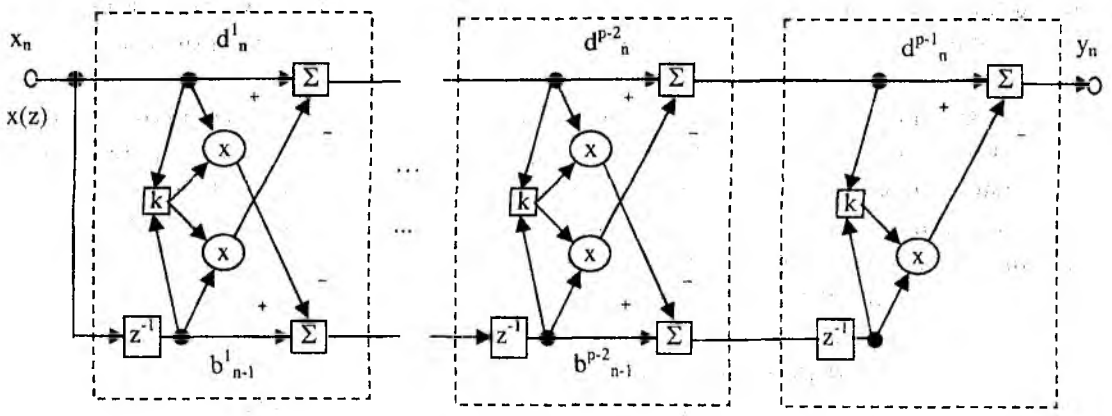

Рис.1

Отметим, что коэффициенты отражения $\vec{k} = (k_1, k_2, \ldots, k_p)$ эффективно вычисляются методом Левинсона.

Коэффициенты отражения $\vec{k} = (k_1, k_2, \ldots, k_p)$ выбеливающего фильтра вычисляют с использованием выборок речевого сигнала, взятых в период молчания.

Декорреляция речевого сигнала выполняется в соответствии с выражениями (3-5), что поясняет структура решетчатого выбеливающего фильтра, показанная на рис. 2.

Алгоритмы декорреляции речевого сигнала в частотной области подробно описаны в литературе [4].

Рассмотрим распознавание речевого сигнала, полученного в результате предварительной обработки.

Для обучающих выборок эталонов производится оценивание параметров $\vec{k}_{em}{}^{(j)} = (k_{em_1}{}^{(j)}, a_{em_2}{}^{(j)}, \ldots, a_{em_p}{}^{(j)})$ обеляющего фильтра для каждого v-го блока речевого сигнала.

Начальные ошибки прямого $d_e{}^{0,j}{}_n$ и обратного предсказания $b_e{}^{0,j}{}_n$ имеют вид

$$d_e{}^{0,j}{}_n = y_{en}{}^j, \quad b^{0,j}{}_n = y_{en}{}^j.$$

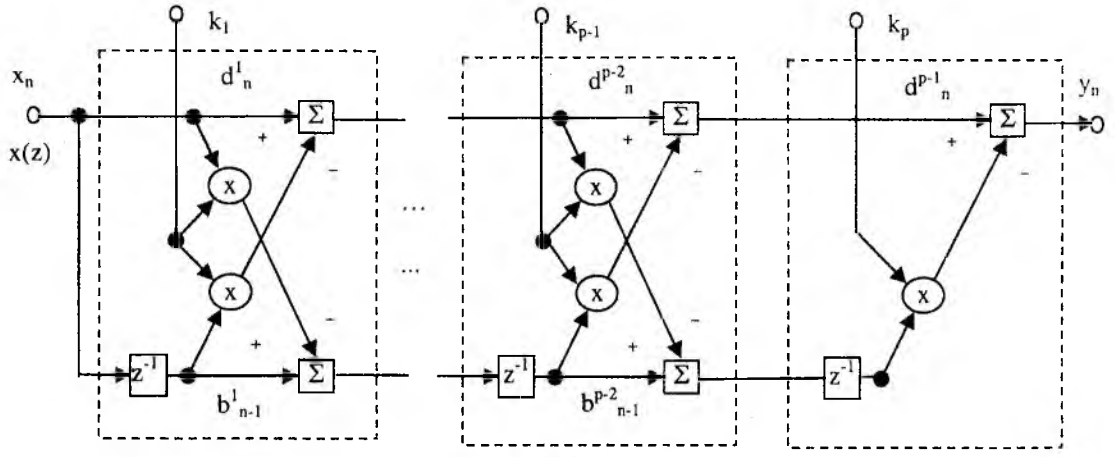

Рис.2

Ошибки прямого $d_e{}^{m,j}{}_n$ и обратного предсказания $b_e{}^{m,j}{}_n$ могут быть представлены в виде

$$d_e{}^{m,j}{}_n = d_e{}^{m-1,j}{}_n - k_m{}^j \cdot b_e{}^{m-1,j}{}_{n-1},$$

$$b_e{}^{m,j}{}_n = b_e{}^{m-1,j}{}_{n-1} - k_m{}^j \cdot d_e{}^{m-1,j}{}_n.$$

Коэффициенты отражения (коэффициенты частной корреляции) могут вычисляться через ошибки предсказания $d^m{}_n$, $b^m{}_n$ в соответствии с выражением

$$k_{em}{}^j = \frac{\sum_{n=1}^{N}(d_e{}^{m-1,j}{}_n \cdot b_e{}^{m-1,j}{}_{n-1})}{(\sum_{n=1}^{N}(d_e{}^{m-1,j}{}_n)^2 \cdot \sum_{n=1}^{N}(b_e{}^{m-1,j}{}_{n-1})^2)^{1/2}},$$

где $m = 1,\ldots, p$.

Коэффициенты отражения согласно методу максимальной энтропии Бурга

$$k_{em}{}^j = \frac{2 \cdot \sum_{n=1}^{N}(d_e{}^{m-1,j}{}_n \cdot b_e{}^{m-1,j}{}_{n-1})}{\sum_{n=1}^{N}(d_e{}^{m-1,j}{}_n)^2 + \sum_{n=1}^{N}(b_e{}^{m-1,j}{}_{n-1})^2}.$$

На этапе принятия решений для каждого блока производится нормирование распознаваемого речевого сигнала

$$y_{t,u}{}^{(k)} = y_{t,u}{}^{(k)} / (\sum_{\tau=1}^{T} y_{\tau,u}{}^{(k)2} / T)^{1/2}. \qquad (6)$$

Рассмотрим алгоритм оценивания результата фильтрации на этапе распознавания.

Полагается, что начальные ошибки прямого $d^0{}_t$ и обратного предсказания $b^0{}_t$ имеют вид

$$d^0{}_{t,u,v}{}^{(k,j)} = y_{t,u}{}^{(k)} \quad \text{и} \quad b^0{}_{t,u,v}{}^{(k,j)} = y_{t,u}{}^{(k)}. \qquad (7)$$

Ошибки прямого $d^{m}{}_{t,u,v}{}^{(k,j)}$ и обратного предсказания $b^{m}{}_{t,u,v}{}^{(k,j)}$ описываются двумя разностными уравнениями

$$d^{m}{}_{t,u,v}{}^{(k,j)} = d^{m-1}{}_{t,u,v}{}^{(k,j)} - k_{m,v}{}^{(j)} \cdot b^{m-1}{}_{t-1,u,v}{}^{(k,j)},$$

$$b^{m}{}_{t,u,v}{}^{(k,j)} = b^{m-1}{}_{t-1,u,v}{}^{(k,j)} - k_{m,v}{}^{(j)} \cdot d^{m-1}{}_{t,u,v}{}^{(k,j)}. \qquad (8)$$

Результат работы выбеливающего фильтра имеет вид

$$n_{t,u,v}{}^{(k,j)} = d^{P}{}_{t,u,v}{}^{(k,j)}. \qquad (9)$$

Решение для k-го сегмента о наличии заданной речевой единицы выносится в соответствии с выражением

$$i(k) = \arg \max_{j \in [1,M]} R^{(k,j)}, \qquad (10)$$

где усредненная мера

$$R^{(k,j)} = \sum_{m=1}^{M} \sum_{v=1}^{V} \sum_{h=H_1}^{H_2} 1 / D(n_{t,v+h,v}{}^{(k,j,m)}), \qquad (11)$$

где $D(n_{t,v+h,v}{}^{(k,j,m)}) = \sum_{t=1}^{N} n_{t,v+h,v}{}^{(k,j,m)^2} / N - (\sum_{t=1}^{N} n_{t,v+h,v}{}^{(k,j,m)} / N)^2$ – оценка дисперсии сигнала с выхода решетчатого фильтра для одного блока, М – количество эталонов.

Рассмотрим решение задачи распознавания речи с использованием процедур динамического программирования (ДП). Процедура ДП служит для нелинейной временной нормализации.

Решение принимается по максимуму меры сходства эталона и текущей распознаваемой речевой единицы. В качестве меры расхождения между векторами признаков вводится расстояние между ними, например в виде

$$d^{(k,n)}(m,v) = 1 / \sum_{t=1}^{N} n_{t,v}{}^{(k,n,m)^2}. \qquad (12)$$

Минимальное значение $D(A)$ достигается при оптимальном согласовании временных расхождений между образами.

Вычисление начинается от концов сегментов слов и завершается к их началу. Рекуррентное уравнение алгоритма представляется следующим образом:

$$D^{(k,n)}(i,j) = \max \begin{cases} D^{(k,n)}(i-1, j-2) + 2d^{(k,n)}(i, j-1) + d^{(k,n)}(i,j), \\ D^{(k,n)}(i-1, j-1) + 2d^{(k,n)}(i,j), \\ D^{(k,n)}(i-2, j-1) + 2d^{(k,n)}(i-1, j) + d^{(k,n)}(i,j). \end{cases} \qquad (13)$$

Альтернативный рекуррентный алгоритм вычисления уравнения представляется следующим образом:

$$D^{(k,n)}(i,j) = \max \begin{cases} D^{(k,n)}(i, j-1) + \chi \cdot d^{(k,n)}(i,j), \\ D^{(k,n)}(i-1, j-1) + d^{(k,n)}(i,j), \\ D^{(k,n)}(i-1, j) + \chi \cdot d^{(k,n)}(i,j). \end{cases} \qquad (14)$$

Решение для к-го сегмента о наличии заданной речевой единицы принимается по правилу

$$i(k) = \arg \max_{n \in [1,M]} D^{(k,n)}(0,0). \qquad (15)$$

Рассмотрим другие алгоритмы распознавания, использующие решетчатые выбеливающие фильтры. Принятие решение для k-го сегмента о слове выносится в соответствии с выражением

$$i(k) = \arg \max_{j \in [1,M]} p(k \mid j), \qquad (16)$$

где условная вероятность

$$p(k \mid j) = \sum_{m=1}^{M} \prod_{v=1}^{V} \sum_{h=H_1}^{H_2} \frac{1}{\sqrt{2\pi\sigma_v}} \exp(-\sum_{t=1}^{N} n_{t,v+h,v}{}^{(k,j,m)^2} / (2\sigma_v{}^2)). \qquad (17)$$

Здесь M – количество эталонов.

Оценки формантных частот спектрально-полосным методом могут вычисляться, как среднеэффективные частоты в заданных полосах частот с соответствующего выхода полосового фильтра с заданными граничными частотами $f_в^{(m)}$ и $f_н^{(m)}$, указанными в табл.1 для каждого из блоков.

Таблица 1

m	$f_н^{(m)}$, Гц	$f_в^{(m)}$, Гц
1	200	850
2	850	2200
3	2200	3000
4	3000	4000

Рассмотрим особенности формирования формантно-полосных признаков. Согласно этому методу вычисляют спектрально-полосные сигналы, соответствующие вероятному расположению формант, полосы которых приведены в табл. 1. Граничные частоты $f_в^{(m)}$, $f_н^{(m)}$ соответствуют m-м формантам при частоте дискретизации 8 кГц .

При этом оценки формантных частот вычисляются путем подсчета количества ноль-пересечений речевого сигнала с соответствующего выхода полосового фильтра с заданными граничными частотами $f_в^{(m)}$ и $f_н^{(m)}$, указанными в табл.1 для каждого из блоков (отрезков) речи, которые берутся с 2-3-х кратным перекрытием или без него.

Улучшить точность первичного оценивания траектории формант можно путем выполнения операции сглаживания $\hat{f}_{ср}^{(m)} = \sum_{r=-v}^{u} \hat{f}^{(m-r)} * W_r$, где $\sum_{r=-v}^{u} W_r = 1$ (экспериментально получено, что такое сглаживание эффективно, когда m>2).

Процедура вычисления формант может быть повторена, но при этом в качестве граничных полос частот используют $\hat{f}_в^{(m)} = \hat{f}^{(m)} + \Delta$, $\hat{f}_н^{(m)} = \hat{f}^{(m)} - \Delta$, где $\hat{f}^{(m)}$ – форманты, вычисленные на предыдущем этапе, Δ – границы диапазона поиска формант. Простейшей среди рекуррентных процедур является двухэтапная.

Расстояние на основе спектрально-полосных оценок для одного эталона на каждое слово имеет вид

$$D_{v,u} = \sum_{h=-J}^{J} \sum_{j=0}^{N_{бл}^{v}-1} D_{v,u}(h,j)/(N_{бл}^{v}), \qquad (18)$$

где $N_{бл}^{v}$ – количество блоков v-го слова

Локальное расстояние может вычисляться как

$$D_{v,u}(h,j) = \sum_{m=1}^{L} | \hat{f}_{j+h}^{(m,u)} - \hat{f}_j^{ob(m,v)} |^r, \qquad (19)$$

где L – количество полос частот (табл.1), v- номер эталона слова, u-номер сегмента.

Логарифмическая мера имеет вид

$$D_{v,u}(h,j) = \log_a(\sum_{m=1}^{L} | \hat{f}_{j+h}^{(m,u)} - \hat{f}_j^{ob(m,v)} |). \qquad (20)$$

Признак вокализованности N_j^u вычисляется путем подсчета количества ноль-пересечений для каждой из выборок j-й выборки u-го сегмента слова.

Определим функционал для признака вокализованности с одним эталоном на каждое слово в виде

$$D_{Вок\,v,u} = \sum_{h=-J}^{J} \sum_{j=0}^{N_{бл}^{v}-1} |N_{j+h}^{(u)} - N_{j}^{ob(v)}|^{r} / (N_{бл}^{v}). \qquad (21)$$

Решение о слове для к-го сегмента на основе авторегрессионых и спектрально-полосных оценок формантных частот, признака вокализованности речи принимается из условия

$$i(k) = \arg\max_{l=\overline{0,M}} \Big(\sum_{d=1}^{D} \sum_{s=0}^{S(d)} (k_{выб} \cdot (R_{выб}^{(k,l,s,d)}) / (\max_{k,l} R_{выб}^{(k,l,s,d)} - \min_{k,l} R_{выб}^{(k,l,s,d)}) -$$
$$- k_{SP} \cdot D_{SP\,l,s}^{(k,d)} / (\max_{k,l} D_{SP\,l,s}^{(k,d)} - \min_{k,l} D_{SP\,l,s}^{(k,d)})) -$$
$$- k_{Вок} \cdot D_{Вок\,l,s}^{(k,d)} / (\max_{k,l} D_{Вок\,l,s}^{(k,d)} - \min_{k,l} D_{Вок\,l,s}^{(k,d)})), \qquad (22)$$

где $R_{выб}^{(k,l,s,d)}$ – функционал, построенный на основе выбеливания речи; $R_{SP}^{(k,l,s,d)}$ – функционал, построенный на основе спектрально-полосных оценок, S(d) – количество эталонов диктора d.

Коэффициенты отражения связаны с площадями A_i сечений голосового тракта

$$k_i = (A_{i+1} - A_i)/(A_{i+1} + A_i), \qquad (23)$$

откуда легко получать, что нормированные площади сечений голосового тракта

$$S_{i+1} = \frac{A_{i+1}}{A_1} = \prod_{j=1}^{i} \frac{1+k_j}{1-k_j}. \qquad (24)$$

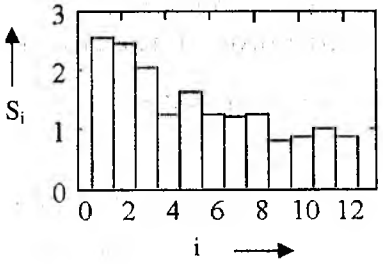

Рис. 3

Нормированные площади сечений голосового тракта показаны на рис.3.

Логарифм отношений площадей сечений голосового тракта

$$\ln(S_{i+1}) = \ln(\frac{A_{i+1}}{A_1}) = \sum_{j=1}^{i} \ln \frac{1+k_j}{1-k_j}. \qquad (25)$$

Принятие решение о наличии заданного слова принимается по правилу

$$i(k) = \arg\min_{v\in[1,M]} D_{v,u}.$$

Определим минимальное расстояние при возможных временных сдвигах как

$$D_{v,u} = \sum_{s=1}^{S} \min_{h=-J,\ldots,J} \sum_{j=0}^{N_{бл}^{v,s}-1} D_{v,u}{}^{s}(h,j)/(N_{бл}^{v,s}), \qquad (26)$$

где S – количество эталонов на одно слово речи, $N_{бл}^{v,s}$ – количество блоков в s-м эталоне v-го слова.

Расстояние как средняя мера для одного эталона имеет вид

$$D_{v,u} = (\sum_{s=1}^{S} \sum_{h=-J}^{J} \sum_{j=0}^{N_{бл}^{v,s}-1} D_{v,u}{}^{s}(h,j))/(N_{бл}^{v,s}), \qquad (27)$$

где локальное расстояние может вычисляться с использованием логарифма отношений площадей сечений голосового тракта

$$D_{v,u}(h,j)=\sum_{m=1}^{p}|\ln(S_{m,j+h}^{(u)})-\ln(S_{m,j}^{ob(v)})|^r \qquad (28)$$

или коэффициентами отражения $\vec{k}=(k_1,k_2,....,k_p)$

$$D_{v,u}(h,j)=\sum_{m=1}^{p}|k_{m,j+h}^{(u)}-k_{m,j}^{ob(v)}|^r . \qquad (29)$$

Исследования для мер r = 1, r = 2, r = 1/2 показали наилучшие результаты для случая r = 1/2.

Для проверки предложенных алгоритмов распознавания речи были проведены экспериментальные исследования. Испытания приведенных выше алгоритмов распознавания слов речи проводились на основе данных, веденных в ЭВМ с микрофона через звуковой интерфейс с частотой дискретизации Fд = 8кГц. В качестве слов речи использовалось десять цифр. Качество распознавания сигналов оценивалось средней вероятностью правильного распознавания, которая получалась на контрольных выборках реализаций методом статистических испытаний.

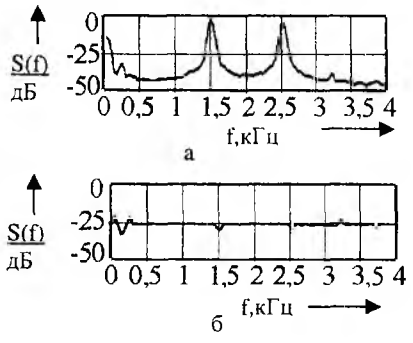

Рис. 4

Предложенный алгоритм (6-11) на основе выбеливающих фильтров позволил получать среднюю вероятности правильного распознавания слов речи 0,91. Для алгоритмов распознавания слов речи с использованием процедуры динамического программирования в соответствии с формулами (13,15) и (14,15) получена средняя вероятность правильного распознавания слов P = 0,91.

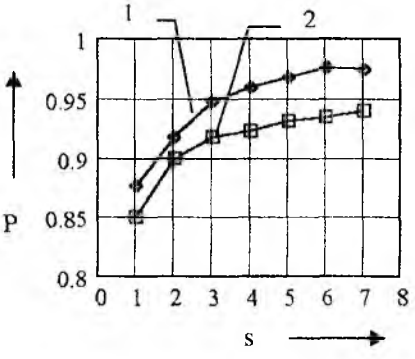

Рис.5

С помощью цифровых фильтров второго порядка из гауссова белого шума получены выборки узкополосного процесса с центральными частотами 1,5 кГц и 2,5 кГц и полосами частот 100 Гц. На рис. 4а показаны энергетические спектры смеси узкополосных случайных процессов, а рис. 4б – энергетические спектры результата подавления помехи фильтром. Алгоритм (3-11) позволил получать при действии смеси двух узкополосных помех с соотношением сигнал-шум по мощности $q^2=1$ среднюю вероятность правильного распознавания слов речи P = 0,5, а после обработки решетчатым фильтром двадцатого порядка (рис. 4б) – среднюю вероятность правильного распознавания слов P = 0,88.

Экспериментально получено, что при принятии решения с одновременным использованием спектрально-полосных признаков, признаков вокализованности и дисперсии результатов выбеливания в алгоритме (22) средняя вероятность правильного распознавания слов P = 0,96.

Для алгоритмов распознавания слов речи по оценкам коэффициентов отражения и использовании одного эталона на каждое из слов получена средняя вероятность правильного распознавания 0,875, а шести эталонах на каждое слово - средняя вероятность правильного распознавания 0,975 (кривая 1 рис. 5).

Зависимость вероятности правильного распознавания для алгоритмов распознавания слов речи по признакам логарифмов отношения площадей сечений голосового тракта от количества используемых s эталонов на каждое слово показана на кривой 2 рис. 5.

Выводы. Таким образом, разработаны алгоритмы распознавания слов речи на основе результата выбеливания решетчатых фильтров, коэффициентов отражения и логарифмов отношения площадей сечений голосового тракта. По найденным рабочим характеристикам проведены сравнительные исследования алгоритмов распознавания слов речи. Проведенные

исследования алгоритмов распознавания подтверждают возможность получения приемлемого качества распознавания речевых сигналов в условиях действия коррелированных помех.

Список литературы: 1. *Пресняков И.Н., Омельченко А.В., Омельченко С.В.* Автоматическое распознавание речи в каналах передачи // Радиоэлектроника и информатика. 2002. №1.С. 26 – 31. 2. *Пресняков И.Н., Омельченко С.В.* Автоматическое распознавание раздельных слов и фонем речи // Там же. 2003. №4.С. 41 – 47. 3. *Пресняков И.Н., Омельченко С.В.* Помехоустойчивые алгоритмы сегментации речи в системах обработки// Радиотехника: Всеукр. межвед. науч.-техн. сб. 2003. №131.С. 165 – 177. 4. *Пресняков И.Н., Омельченко С.В.* Алгоритмы распознавания фонем речи// Там же. 2003. №135.С. 180 – 189. 5. *Рабинер Л. Р., Шафер Р.В.* Цифровая обработка речевых сигналов: Пер. с англ. / Под ред. М. В. Назарова и Ю. Н. Прохорова. М.: Радио и связь, 1981. 496 с. 6. *Марпл.-мл. С.Л.* Цифровой спектральный анализ и его приложения: Пер. с англ. М.: Мир, 1990. 584 с. 7. *Дж. Д. Маркел, А. Х. Грей.* Линейное предсказание речи. М.: Связь, 1980. 308с.

Харьковский национальный
университет радиоэлектроники

Поступила в редколлегию 27.02.2004