УДК 534.78: 621.39: 681.513.6

ИССЛЕДОВАНИЕ АЛГОРИТМА КАЛМАНОВСКОЙ ФИЛЬТРАЦИИ РЕЧЕВОГО СИГНАЛА, НАБЛЮДАЕМОГО В ШУМЕ

В.Г. Санников, профессор МТУСИ, к.т.н.; tes_mtuci@mail.ru

Ключевые слова: *речевой сигнал, шум, идентификация, показатель затухания, оптимальная линейная фильтрация, отношение сигнал/шум.*

Введение. В системе речевой связи, простейшая схема которой приведена на рис. 1, речевой сигнал (РС) $x(t)$ формируется под действием сигнала $u(t)$ источника голосового возбуждения (ИГВ) на выходе голосового тракта (ГТ) модели речеобразования [1–3]. При этом наблюдению подлежит не сам РС, а его зашумленная копия $r(t)=x(t)+n(t)$, где $n(t)$ – реализация шума передачи.

В случае, когда уровень шума больше или равен уровню РС, а отношение сигнал/шум ($ОСШ_{вх}$) меньше или равно 0 дБ (мобильная телефония), качество восприятия РС по показателю разборчивости резко ухудшается. Для улучшения отношения сигнал/шум ($ОСШ_{вых}$) на выходе оптимального линейного фильтра (ОЛФ) по сравнению с $ОСШ_{вх}$ и, соответственно, для повышения качества восприятия речи актуальной становиться задача оптимальной фильтрации наблюдаемой смеси $r(t)$.

Известны примеры решения задачи оптимальной линейной фильтрации применительно к обработке зашумленной речи [1–3]. Однако большинство из них основаны на представлении РС $x(t)$ моделью авторегрессии, ограниченной полюсной моделью ГТ, а рекуррентная оценка параметров авторегрессии осуществляется на основе метода наименьших квадратов (МНК), обладающего бесконечной памятью. Такая оценка справедлива для случайных сигналов, относящихся к классу стационарных.

Речевые сигналы входят в класс нестационарных случайных сигналов. Поэтому для точной оценки динамики изменения во времени статистики более адекватен МНК с взвешиванием (с конечной памятью) [4], характеризуемый «множителем забывания» v. Кроме того, в [1] указывается, что более адекватной моделью голосового тракта является его нуль-полюсная модель или модель РС, построенная на основе авторегрессии и скользящего среднего (АРСС модель РС). Наиболее заметно это при анализе и синтезе назализованных звуков речи.

Поскольку в литературе вопросам фильтрации зашумленной речи, учитывающим отмеченные выше особенности модели речеобразования, уделено незначительное внимание, актуально решение следующих задач:

- теоретическое и экспериментальное исследование ОЛФ зашумленной речи с учетом нуль-полюсной или АРСС модели РС;
- оценка параметров ОЛФ на основе МНК с конечной памятью;

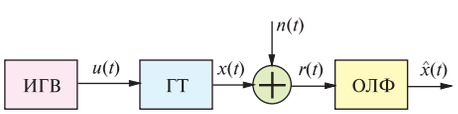

Рис. 1

- оптимизация «множителя забывания» v_{opt}, обеспечивающего экстремум показателю качества оптимальной фильтрации зашумленной речи.

При реализации алгоритмов фильтрации сигналов часто применяют цифровые методы, а сигналы рассматриваются в дискретном времени. Оптимальная фильтрация предполагает наличие априорных сведений о структуре ГТ модели речеобразования. Поэтому сначала определим АРСС модель РС.

Идентификация АРСС модели РС. Нуль-полюсная или АРСС модель РС определяется разностным уравнением

$$x_t = \sum_{i=1}^{m} a_{i,t} x_{t-i} + \sum_{j=0}^{m-1} b_{j,t} u_{t-j}, \qquad (1)$$

показывающим РС как отклик $x_t, t = 0,1,2,...$, устойчивой нестационарной динамической линейной системы, характеризующей ГТ модели речеобразования (рис. 1). На ее вход действует случайная последовательность $u_t, t = 0,1,2,...$, вырабатываемая ИГВ.

Идентификация (в узком смысле) представляет собой процесс определения вектора параметров модели РС (1) по результатам измерения значений $\{x_t\}, t = 0,1,2,...$. Пусть параметры $\{a_{i,t}\}, i = \overline{1,m}$, и $\{b_{j,t}\}, j = \overline{0,m-1}$, модели (1) неизвестны.

Введем обобщенный вектор параметров

$$\begin{aligned}\mathbf{a}_t &= [a_{1,t},...,a_{m,t},b_{0,t},...,b_{m-1,t}]^T = \\ &= [a_{1,t},...,a_{m,t},a_{m+1,t},a_{m+2,t},...,a_{2m,t}]^T\end{aligned} \qquad (2)$$

и расширенный вектор результатов измерений

$$\mathbf{s}_t = [x_{t-1},...,x_{t-m},u_t,...,u_{t-m+1}]^T, \qquad (3)$$

где T – знак векторного транспонирования.

Для оценки искомого вектора параметров (2) модели (1) воспользуемся рекуррентным алгоритмом МНК с взвешиванием, при котором минимизации подвергается взвешенная сумма квадратов «невязки» [5]:

$$\overline{\sigma_t^2}(v) = \sum_{n=0}^{t} v^{t-k}(x_n - \mathbf{s}_n^T \mathbf{a}_n)^2, \qquad (4)$$

где $v \le 1$ – «множитель забывания», характеризующий конечную память алгоритма МНК.

В результате последовательная (рекуррентная) оценка $\hat{\mathbf{a}}_n, n = 0,1,2,...$, вектора параметров (2) АРСС модели РС (1) осуществляется следующим образом [4]:

$$\hat{\mathbf{a}}_n = \hat{\mathbf{a}}_{n-1} + \mathbf{k}_{\mathbf{a},n}(x_n - \hat{\mathbf{s}}_{n-1}^T \hat{\mathbf{a}}_{n-1}), \quad n = 1,2,...; \qquad (5)$$

$$\mathbf{k}_{\mathbf{a},n} = \mathbf{P}_{n-1}\hat{\mathbf{s}}_{n-1} / (v + c_n), \quad c_n = \hat{\mathbf{s}}_{n-1}^T \mathbf{P}_{n-1}\hat{\mathbf{s}}_{n-1}; \qquad (6)$$

$$\mathbf{P}_n = (\mathbf{E}_{2m-1} - \mathbf{k}_{\mathbf{a},n}\hat{\mathbf{s}}_{n-1}^T)\mathbf{P}_{n-1} / v \qquad (7)$$

с начальными условиями: $b_{0,0} = a_{m+1,0} = 1, \mathbf{a}_0 = [0,...,0,1,0,...,0]^T$, $\mathbf{P}_0 = p\mathbf{E}_{2m-1}$. Здесь $\mathbf{E}_{2m-1}$ – единичная $(2m-1)\times(2m-1)$ матрица; p – достаточно большое положительное число (выбрано

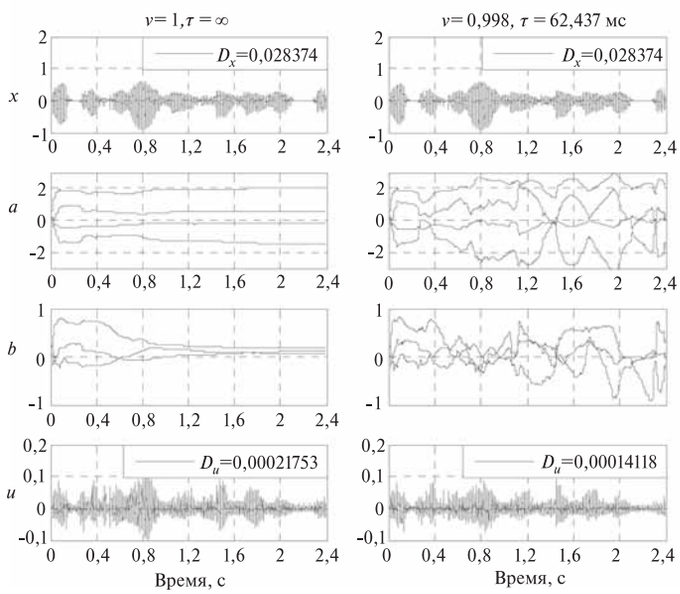

Рис. 2

p=200); $\mathbf{k}_{\mathrm{a},n}$ – вектор коррекции сигнала голосового возбуждения; $\mathbf{P}_n = \mathbf{R}_{\mathrm{s},n}^{-1}$ – матрица размера $(2m-1)\times(2m-1)$, обратная корреляционной матрице $\mathbf{R}_{\mathrm{s},n} = \mathrm{M}\{\hat{\mathbf{s}}_n\hat{\mathbf{s}}_n^T\}$ случайного вектора $\hat{\mathbf{s}}_n$; M – знак математического ожидания.

Заметим, что вектор $\mathbf{s}_t$ в (3) содержит ненаблюдаемые величины $u_{t-1},...,u_{t-m+1}$. Их оценки определяют по оценкам параметров АРСС:

$$\hat{u}_t = (x_t - \hat{\mathbf{s}}_{t-1}^T\hat{\mathbf{a}}_t), \tag{8}$$

где $\hat{\mathbf{s}}_t = [x_{t-1},...,x_{t-m},\hat{u}_t,...,\hat{u}_{t-m+1}]$.

Результаты экспериментальной проверки алгоритма идентификации АРСС модели реального РС приведены на рис. 2. В качестве фонетически сбалансированной обрабатывалась стандартная фраза [6]: «*Эти жирные сазаны ушли под палубу*» (верхние графики) с дисперсией D_x=0,28374. Интервал дискретизации выбирался равным $\Delta t = 125$ мкс. Графики АР $\{a_{i,t}\}, i=\overline{1,4}$, и СС $\{b_{j,t}\}, j=\overline{1,3}$, параметров, а также оценки $\hat{u}_t$ сигнала ИГВ изображены в левой и правой колонках с различными «множителями забывания»: 1) v=1 с интервалом памяти МНК $\tau\to\infty$ (левая колонка); 2) v=0,998 с τ=62,437 мкс (правая колонка). Здесь $v = \exp(-\Delta t/\tau)$, $\tau = \Delta t/\ln(1/v)$.

Анализ графиков рис. 2 позволяет сделать следующие выводы. При v=1 параметры АРСС модели РС существенно изменяются только на незначительном временном интервале, после которого они практически постоянны и не отслеживают изменение статистики РС. При v=0,998 динамика параметров существенно изменяется: они более детально описывают изменение статистики РС, приводя к уменьшению дисперсии сигнала ИГВ от D_u=0,00021753 при v=1 до D_u=0,00014118 при v=0,998.

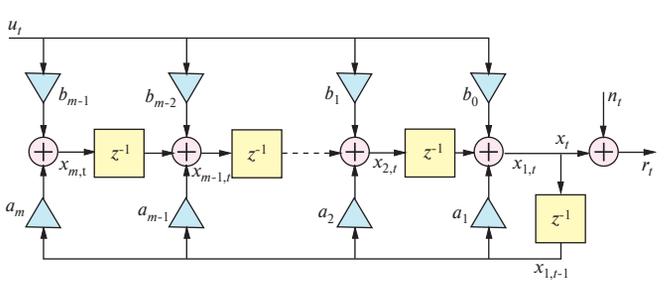

Рис. 3

Для конструктивного использования оптимальной калмановской фильтрации зашумленной речи, определяемой векторно-матричными уравнениями состояния и наблюдения, необходимо от скалярной модели ГТ (1) перейти к ее векторно-матричному представлению.

Уравнения состояния и наблюдения ГТ. Учитывая, что параметры АРСС модели РС можно считать постоянными на интервале длительностью 20 мс [1], построим векторную модель РС в пространстве состояний. Для этого введем переменные состояния ГТ $\{x_{k,t}\}, k=\overline{1,m}$, следующим образом:

$$\begin{cases} x_{m,t} = (a_m x_{t-1} + b_{m-1}u_t); \\ x_{m-1,t} = (a_{m-1}x_{t-1} + b_{m-2}u_t) + a_m x_{t-2} + b_{m-1}u_{t-1} = \\ \quad = (a_{m-1}x_{t-1} + b_{m-2}u_t) + x_{m,t-1}; \\ x_{m-2,t} = (a_{m-2}x_{t-1} + b_{m-3}u_t) + a_{m-1}x_{t-2} + b_{m-2}u_{t-1} + a_m x_{t-3} + \\ \quad + b_{m-1}u_{t-2} = (a_{m-2}x_{t-1} + b_{m-3}u_t) + x_{m-1,t-1}; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_{1,t} = (a_1 x_{t-1} + b_0 u_t) + a_2 x_{t-2} + b_1 u_{t-1} + \ldots + a_m x_{t-m} + \\ \quad + b_{m-1}u_{t-m+1} = (a_1 x_{t-1} + b_0 u_t) + x_{2,t-1}, \quad b_0 = 1. \end{cases} \tag{9}$$

Сравнивая последнюю строку в (9) и соотношение (1), заключаем, что $x_t \equiv x_{1,t}$, $x_{t-1} \equiv x_{1,t-1}$. Теперь, с учетом шумов наблюдения, модель ГТ представляется:

- уравнением состояния

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{b}u_t, \tag{10}$$

- уравнением наблюдения

$$r_t = \mathbf{h}^T\mathbf{x}_t + n_t = x_t + n_t, \tag{11}$$

где $\mathbf{x}_t = [x_{1,t}, x_{2,t},..., x_{m,t}]^T$ – вектор состояния ГТ в момент t;

$$\mathbf{A} = \begin{bmatrix} a_1 & 1 & 0 & \ldots & 0 \\ a_2 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m-1} & 0 & 0 & \ldots & 1 \\ a_m & 0 & 0 & \ldots & 0 \end{bmatrix}$$

– переходная матрица ГТ размера $m\times m$, составленная из параметров авторегрессии (АР) модели (1); $\mathbf{b} = [1, b_1,..., b_{m-1}]^T$ – вектор параметров скользящего среднего (СС) модели (1); $\mathbf{h} = [1,0,0,...,0]^T$ – вектор наблюдения РС; n_t – гауссовская шумовая последовательность с корреляционной функцией $R_{n,t} = \sigma_{n,t}^2\delta_{t,\tau}$; $\delta_{t,\tau}$ – символ Кронекера, n_t и u_t некоррелированы.

Структурная схема ГТ, соответствующая уравнениям состояния (10) и наблюдения (11), показана на рис. 3.

Оптимальная линейная фильтрация РС. Задача фильтрации состоит в том, чтобы по совокупности последовательных измерений $\{r_1, r_2,..., r_t\} = \{r_k\}, k=\overline{1,t}$, найти оптимальную оценку $\hat{\mathbf{x}}_t$ вектора состояния $\mathbf{x}_t$, удовлетворяющую критерию минимума текущей среднеквадратической погрешности фильтрации $\sigma_t^2 = \mathrm{M}\{\mathbf{e}_t^T\mathbf{e}_t\}$, где $\mathbf{e}_t = (\mathbf{x}_t - \hat{\mathbf{x}}_t)$.

Для синтеза ОЛФ, в соответствии с полученной моделью ГТ, можно воспользоваться хорошо разработанной теорией ОЛФ в пространстве состояний, приводящей к фильтру Калмана, работа которого определяется соотношениями [4]:

$$\begin{aligned} \hat{\mathbf{x}}_t &= \mathbf{A}_{t-1}\hat{\mathbf{x}}_{t-1} + \mathbf{k}_{\mathrm{x},t}[r_t - \mathbf{h}^T\mathbf{A}_{t-1}\hat{\mathbf{x}}_{t-1}] = \\ &= \mathbf{A}_{t-1}\hat{\mathbf{x}}_{t-1} + \mathbf{k}_{\mathrm{x},t}[r_t - (a_{1,t-1}\hat{x}_{1,t-1} + \hat{x}_{2,t-1})], t = 1,2,...; \end{aligned} \tag{12}$$

$$\begin{aligned} \mathbf{k}_{\mathrm{x},t} &= \mathbf{P}_{t|t-1}\mathbf{h}(R_{n,t} + \mathbf{h}^T\mathbf{P}_{t|t-1}\mathbf{h})^{-1} = \\ &= \mathbf{V}_t\mathbf{h}R_{n,t}^{-1} = [V_{11}, V_{21},..., V_{m1}]_t^T / \sigma_{n,t}^2; \end{aligned} \tag{13}$$

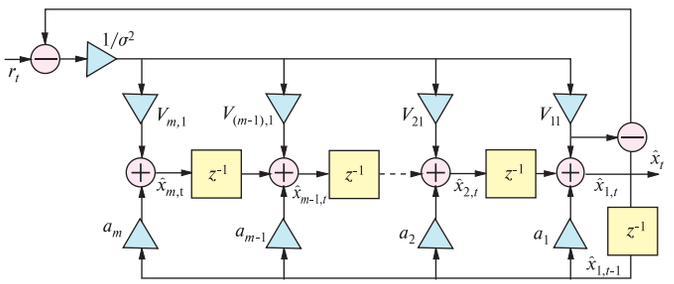

Рис. 4

$$\mathbf{P}_{t|t-1} = \mathbf{A}_{t-1}\mathbf{V}_{t-1}\mathbf{A}_{t-1}^T + \mathbf{b}_{t-1}\mathbf{R}_{u.t-1}\mathbf{b}_{t-1}^T; \qquad (14)$$

$$\mathbf{V}_t = (\mathbf{E}_m - \mathbf{k}_{\mathbf{x},t}\mathbf{h}^T)\mathbf{P}_{t|t-1} \qquad (15)$$

с начальными условиями: $\mathbf{x}_0 = \mathrm{M}\{\mathbf{x}_0\} = \bar{\mathbf{x}}_0$, $\mathbf{V}_0 = \mathrm{M}\{(\mathbf{x}_0 - \bar{\mathbf{x}}_0)(\mathbf{x}_0 - \bar{\mathbf{x}}_0)^T\}$. Здесь $\mathbf{k}_{\mathbf{x},t}$ – векторный коэффициент усиления ОЛФ; $\mathbf{P}_{t|t-1}$ – корреляционная матрица погрешности экстраполяции на один шаг (априорная корреляционная матрица погрешности фильтрации); $\mathbf{V}_t = [V_{11}, V_{21}, ..., V_{m1}]_t^T$ – апостериорная корреляционная матрица погрешности фильтрации; $\mathbf{R}_{u,t}$ – корреляционная матрица сигнала возбуждения ИГВ; $\mathbf{E}_m$ – единичная $m \times m$ матрица; $\sigma_{n,t}^2$ – текущая дисперсия шума наблюдения.

Структурная схема ОЛФ для оценки РС, наблюдаемого на фоне шума, равномерно распределенного в полосе частот сигнала, показана на рис. 4. Здесь полюса системной функции цифрового ОЛФ (нижняя часть схемы) определяются АР параметрами $\{a_j\}$, $j = \overline{1,m}$, модели ГТ, в то время как ее нули характеризуются не СС параметрами $\{b_j\}$, $j = \overline{0,m-1}$, модели ГТ (рис. 3), а параметрами $\{V_{j1}\}$, $j = \overline{1,m}$, первого столбца матрицы погрешности фильтрации (верхняя часть схемы).

Результаты экспериментальной проверки алгоритма оптимальной линейной (калмановской) фильтрации зашумленного РС представлены рис. 5 и таблицей, в которую сведены величины v_{opt}, $\tau_{opt} = \Delta t / \ln(1/v_{opt})$ и максимальные

m	2	4	6	8	10	12	14
v_{opt}	0,9990	0,9950	0,9930	0,9916	0,9915	0,9915	0,9915
τ_{opt}, мс	124,94	24,937	17,795	14,818	14,643	14,643	14,643
ОСШ$_{вых}$, дБ	5,6768	6,5224	7,1988	7,2769	7,6234	7,4988	7,2981

значения ОСШ$_{вых}$ в зависимости от m. Результаты получены для ОСШ$_{вх}$=0 дБ.

Как и ранее обработке подвергалась зашумленная фраза: «*Эти жирные сазаны ушли под палубу*» с различными ОСШ$_{вх}$ = $=10\lg(\sigma_x^2/\sigma_n^2)$, дБ. Экспериментальные данные получены для различного числа АР и СС параметров модели ГТ, а также для разных значений «множителя забывания» v МНК с взвешиванием. Видно, что при фиксированной памяти моделей ГТ и ОЛФ, определяемой параметром m, значение ОСШ$_{вых}$=$10\lg[\sigma_x^2/(x_t - x_t)^2]$ изменяется так, что при некоторой величине v_{opt} принимает максимальное значение.

Наличие v_{opt} объясняется следующим. При $v \to 1$ память алгоритма МНК бесконечна ($\tau \to \infty$), а корреляционная матрица речи в (7) и параметры предсказания в (5) значительно усреднены и не успевают отслеживать изменение статистики РС. При уменьшении v и τ возрастает вес краткосрочной корреляции ($\mathbf{P}_n$ в (7) увеличивается). Следовательно, корреляционная матрица речи в (7) и параметры предсказания в (5) начинают лучше отслеживать изменение статистики РС. Это приводит к уменьшению погрешности предсказания и фильтрации, а также возрастанию ОСШ$_{вых}$. Однако при дальнейшем уменьшении v и τ дисперсии параметров предсказания возрастают из-за увеличения в них вредной доли сигнала голосового возбуждения, что приводит к ухудшению предсказания и в результате к уменьшению ОСШ оптимальной фильтрации РС.

Графическая иллюстрация результатов эксперимента приведена на рис. 5. Здесь в правой колонке (сверху вниз) показаны фрагменты: исходного РС $\{x_t\}$, зашумленного РС

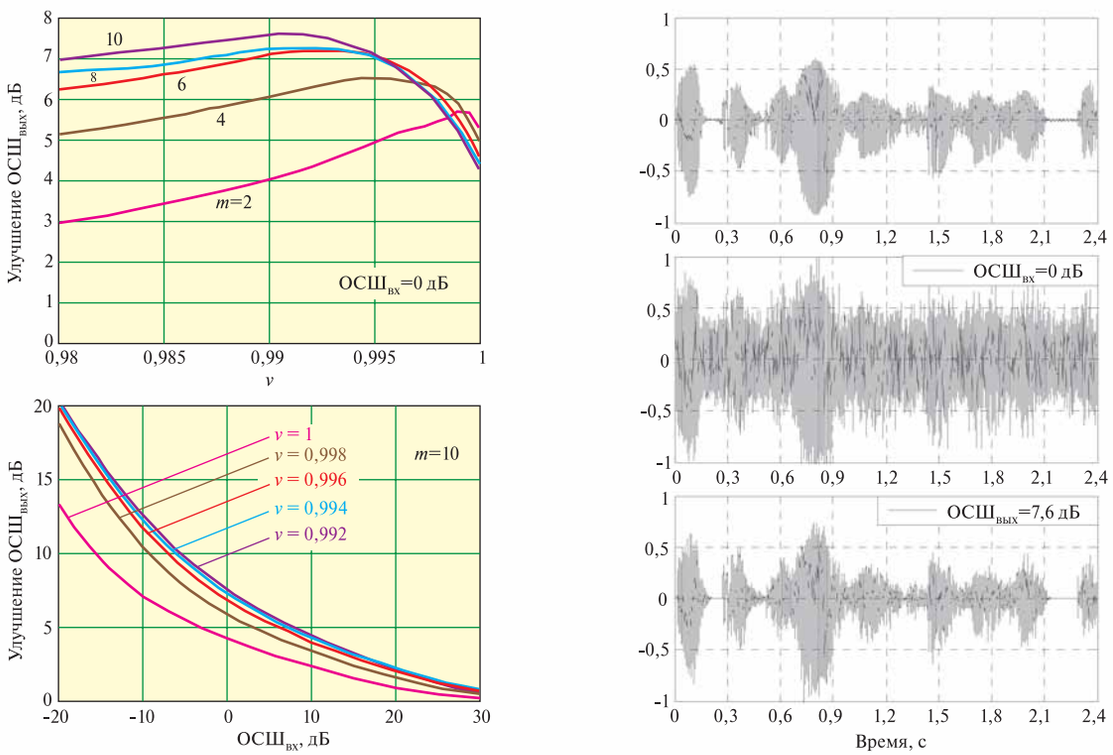

Рис. 5

$\{r_i\}$ и оценки РС $\{x_i\}$, наблюдаемой на выходе ОЛФ. Нетрудно заметить, что фильтр значительно подавляет шум, особенно в паузах речи.

В левой колонке (верхние графики) показаны зависимости $ОСШ_{вых}$ от «множителя забывания» v при различных m и $ОСШ_{вх} = 0$ дБ. Явно виден экстремальный характер этих зависимостей, оптимальные значения которых приведены в таблице.

На нижнем графике показаны кривые улучшения $ОСШ_{вых}$ ($ОСШ_{вых} - ОСШ_{вх}$), обеспечиваемого ОЛФ в зависимости от $ОСШ_{вх}$ при различных v и m=10. Из графиков следует, что при фиксированном $ОСШ_{вх}$ МНК при v=1 уступает по эффективности фильтрации МНК с v<1. Так, при $ОСШ_{вх} = 0$ дБ и $v = 1$ (интервал памяти $\tau \to \infty$) $ОСШ_{вых} = 4{,}4$ дБ, в то время как при v=0,992 (интервал памяти τ=14,85 мс) $ОСШ_{вых}$=7,62 дБ, т.е. на 3,22 дБ больше.

Заключение. Дальнейшее совершенствование методов оптимальной фильтрации зашумленной речи с повышенной помехозащищенностью возможно только при использовании априорных сведений, более адекватных моделям формирования и слухового приема реальных речевых сигналов.

Это направление развивается в лаборатории речевой информатики и связи МТУСИ и представляет значительный интерес для идентификации личности говорящего по голосу и речи, являющейся одной из приоритетных задач, решаемых на базе комбинированной обработки речевых сигналов, повышения разборчивости речи и голосовой биометрии.

ЛИТЕРАТУРА

1. **Маркел Дж. Д., Грэй А.Х.** Линейное предсказание речи: Пер. с англ./ Под ред. Ю.Н. Прохорова и В.С. Звездина. – М.: Связь, 1980. – 308 с.
2. **Прохоров Ю.Н.** Статистические модели и рекуррентное предсказание речевых сигналов. – М.: Радио и связь, 1984. – 240 с.
3. **Шелухин О.И., Лукьянцев Н.Ф.** Цифровая обработка и передача речи/ Под ред. О.И. Шелухина. – М.: Радио и связь, 2000. – 456 с.
4. **Сейдж Э., Мелс Дж.** Теория оценивания и её применение в связи и управлении: Пер. с англ./ Под ред. Б.Р. Левина. – М.: Связь, 1976. – 496 с.
5. **Санников В.Г.** Статистический анализ методов формирования речевых сигналов. – М.: МТУСИ, 2005. – 140 с.
6. ГОСТ Р 50840-95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости.

Получено после доработки 11.03.11