МЕТОД ОПТИМАЛЬНОЙ ФИЛЬТРАЦИИ ЗАШУМЛЕННОЙ РЕЧИ НА ОСНОВЕ ПСИХОАКУСТИЧЕСКОЙ МОДЕЛИ ЕЁ СЛУХОВОГО ВОСПРИЯТИЯ

© 2016 г. В.Г. САННИКОВ

Московский технический университет связи и информатики

Введение. В теории и технике связи проблема помехоустойчивости занимает центральное место. Задача синтеза системы связи может считаться полностью определенной только тогда, когда сформулированы требования получателя к верности воспроизведения переданных сигналов. Это положение особенно справедливо в отношении систем речевой связи, в которых получателем речевых сигналов является слуховая система человека [1], анализирующая поступающие на её вход зашумленные стимулы и осуществляющая распознавание передаваемых сообщений с оценкой их качества по громкости и разборчивости [2,3].

В виду того, что ни одна из известных технических систем передачи речи не может сравняться по производительности и помехоустойчивости с биологической слуховой системой, актуальна задача исследования преобразований речи на периферии слуха и построения их математических моделей. В работах [1,3] показано, что слуховая система воспринимает входной одномерный сигнал по многим сенсорным каналам. Очевидно, использование нескольких каналов при оценивании одного и того же сигнала в шуме и позволяет слуховой системе увеличить точность оценивания и повысить надежность распознавания передаваемых сообщений, содержащихся в речевом сигнале. Принимая гипотезу о том, что в слуховой системе речевой сигнал подвергается оптимальным преобразованиям (это справедливо, по крайней мере, для малых и средних уровней сигнала) [4], рассмотрим задачу многоканальной оптимальной линейной фильтрации речи с учетом известных свойств слуховой системы.

Идентификация модели речеобразования. Для решения задач оптимальной фильтрации зашумленной речи требуются априорные сведения о модели речеобразования. В практике речевой связи часто используют авторегрессионную модель речеобразования [5]. В дискретном времени она принимает вид:

$$s_t = a_{1,t}s_{t-1} + a_{2,t}s_{t-2} + ... + a_{m,t}s_{t-m} + u_t, \quad t = 1,2,..., \tag{1}$$

где $\mathbf{a}_t = (a_{1,t}, a_{2,t}, ..., a_{m,t})^T$, $\mathbf{s}_{t-1} = (s_{t-1}, s_{t-2}, ..., s_{t-m})^T$, соответственно векторы параметров модели и отсчетов речевого сигнала, определенные для некоторого фиксированного момента времени t, m – порядок модели, u_t - случайная последовательность гауссовского шума модели с независимыми значениями, нулевым средним и дисперсией $D_{u,t}$.

Собственно под идентификацией модели (1) понимают определение по значениям чистой речи $s_t, t = 1,2,...,$ её параметров, к которым относятся: $\mathbf{a}_t$ и $D_{u,t}$. При этом последовательная оценка вектора параметров $\mathbf{a}_t$ осуществляется на основе рекуррентного метода наименьших квадратов [6]:

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \mathbf{k}_{\mathbf{a},t}(s_t - \mathbf{s}_{t-1}^T\mathbf{a}_{t-1}), \quad \mathbf{k}_{\mathbf{a},t} = \mathbf{R}_t^{-1}\mathbf{s}_{t-1}, \quad \mathbf{a}_0 = 0, \tag{2}$$

$$\mathbf{R}_t^{-1} = [\mathbf{R}_{t-1}^{-1} - \mathbf{R}_{t-1}^{-1}\mathbf{s}_{t-1}(1 + \mathbf{s}_{t-1}^T\mathbf{R}_{t-1}^{-1}\mathbf{s}_{t-1})^{-1}\mathbf{s}_{t-1}^T\mathbf{R}_{t-1}^{-1}], \quad \mathbf{R}_0^{-1} = 200 \cdot \mathbf{I}, \tag{3}$$

где $\mathbf{R}_t$ - матрица корреляции выборки речевого сигнала, $\mathbf{R}_t^{-1}$ - её обратная матрица, $\mathbf{I}$ - единичная матрица, размера $m \times m$.

В качестве шума модели речеобразования воспользуемся оценкой погрешности предсказания: $u_t = s_t - \mathbf{s}_{t-1}^T \mathbf{a}_t$. Дисперсия этой последовательности определяется из условия минимума среднеквадратичной погрешности предсказания

$$D_{u,t} = D_{s,t} - \mathbf{b}^T \mathbf{R}_t \mathbf{a}_t, \quad \mathbf{b}^T = (0\,0\,...\,1), \tag{4}$$

где $D_{s,t}$ - рекуррентная оценка дисперсии последовательности речевого сигнала

$$D_{s,t} = D_{s,t-1} - [D_{s,t-1} - s_t^2]/t\,. \tag{5}$$

Уравнение состояния модели речеобразования. Перейдем от скалярного представления речевого сигнала в (1) к канонической векторной форме в пространстве состояний. Введем векторы переменных состояния: $\mathbf{x}_{t+1} = (s_{t-m}, \mathrm{s}_{t-m+1}, ..., s_t, s_{t+1})^T$ и $\mathbf{x}_t = (s_{t-m+1}, \mathrm{s}_{t-m+2}, ..., s_t)^T$. Тогда скалярному уравнению (1) будет соответствовать уравнение состояния модели речеобразования в векторно-матричной форме:

$$\mathbf{x}_{t+1} = \mathbf{A}_t \mathbf{x}_t + \mathbf{b} u_t\,, \tag{6}$$

Где, например, при $m = 4$ матрица $\mathbf{A}_t$ и вектор-столбец $\mathbf{b}$ принимают следующий вид

$$\mathbf{A}_t = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a_{4,t} & a_{3,t} & a_{2,t} & a_{1,t} \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{7}$$

Учет свойств слуховой системы. Полагаем, что входной речевой сигнал в смеси с внешним шумом на периферии слуха воздействует на N чувствительных стереоцилий волосковых клеток, колеблющихся с частотами $f_r(z_n), z_n = n\Delta z, n = \overline{1,N}$, где z_n - высота тона в барках [1]. Этот процесс эквивалентен слуховой фильтрации зашумленной речи в различных высотных каналах с коэффициентами передачи [2]:

$$k(z_n, f) = \left\{1 + \left[\left(f_r^2(z_n) \cdot f^{-1} - f\right) / \Delta f(z_n)\right]^2\right\}^{-5}. \tag{8}$$

где $f_r(z_n) = 0.579\mathrm{sh}(0.1654 z_n)$, $\Delta f(z_n) = 0.08788\mathrm{ch}(0.1654 z_n)$ кГц - резонансные частоты и полосы частот слуховых фильтров с импульсными реакциями $h_{n,t}, n = \overline{1,N}$, определяемых как преобразование Фурье от функции $k(z_n, f)$.

Следует также учитывать, что распространение входного стимула в улитке внутреннего уха сопровождается совместным действием внешнего и внутренних сенсорных шумов $v_{n,t}, n = \overline{1,N}$, распределенных неравномерно по разным каналам и вызывающих эффекты маскировки речи, проявляющиеся в наличии порогов слышимости. Так, например, кривые порога слышимости при внешнем белом шуме определяются соотношением [2]:

$$L_v(z_n) = l_{bsh} + 26{,}5 + L_\Delta(z_n) + 10\log\left\{1 + 10^{0{,}1[L_0(z_n) - l_{bsh} - 26{,}5 - L_\Delta(z_n)]}\right\}, \tag{9}$$

где l_{bsh} - уровень внешнего белого шума, $L_\Delta(z_n) = 10\log\Delta f(z_n)$, $L_0(z_n)$ - уровень абсолютного порога слышимости в тишине [1,2].

На основе (9) дисперсии шумов $v_{n,t}, n = \overline{1,N}$, определяются так

$$D_v(z_n) = P_0 \cdot 10^{0{,}1 L_v(z_n)}, \tag{10}$$

где P_0 – величина нулевого порога восприятия для нормального слуха, пропорциональная пороговой интенсивности $I_0 = 10^{-12}$ Вт/м^2.

Уравнения наблюдений. С учетом сказанного выше процесс $\mathbf{x}_t$ в (6) наблюдается многоканальной слуховой системой, состоящей из N каналов:

$$y_{n,t} = \mathbf{h}_{n,t}^T \mathbf{x}_t + v_{n,t}, \quad n = \overline{1,N}. \tag{11}$$

Здесь $\mathbf{h}_{n,t}$ - векторы наблюдений слуховой системы, $v_{n,t}$ - случайные шумы наблюдения с независимыми значениями, нулевыми средними и дисперсиями в (10).

Уравнения оценки вектора состояния модели речеобразования. Для получения оптимальной оценки $\widehat{\mathbf{x}}_t$ вектора состояния $\mathbf{x}_t$ можно воспользоваться уравнениями калмановской фильтрации [6]. С учетом ряда преобразований приходим к уравнению оценки для алгоритма многоканальной фильтрации:

$$\widehat{\mathbf{x}}_t = \mathbf{x}_{p,t} + \sum_{n=1}^{N} \mathbf{k}_{n,t}\varepsilon_{n,t}, \ \mathbf{x}_{p,t} = \mathbf{A}_t \widehat{\mathbf{x}}_{t-1}, \ \varepsilon_{n,t} = \left(y_{n,t} - \mathbf{h}_{n,t}^T \mathbf{x}_{p,t}\right), \ \mathbf{k}_{n,t} = \mathbf{V}_t \mathbf{h}_{n,t} D_{v,n,t}^{-1}, \tag{12}$$

где $\mathbf{x}_{p,t}$ - вектор прогноза на один такт, $\varepsilon_{n,t}$ - «обновляющий» процесс, $\mathbf{k}_{n,t}$ - векторный коэффициент усиления n–ого канала оптимального фильтра, $\mathbf{V}_t = \overline{(\widehat{\mathbf{x}}_t - \mathbf{x}_t)(\widehat{\mathbf{x}}_t - \mathbf{x}_t)^T}$ - корреляционная матрица ошибок оптимальной фильтрации, удовлетворяющая следующему соотношению

$$\mathbf{V}_t^{-1} = \mathbf{P}_{t-1}^{-1} + \sum\nolimits_{n=1}^{N} \mathbf{h}_{n,t} D_{v,n,t}^{-1} \mathbf{h}_{n,t}^T . \tag{13}$$

Здесь $\mathbf{P}_{t-1}$ - корреляционная матрица ошибок экстраполяции, равная

$$\mathbf{P}_{t-1} = \mathbf{A}_t \mathbf{V}_{t-1} \mathbf{A}_t^T + \mathbf{b} D_{\mathrm{u},t} \mathbf{b}^T, \ \mathbf{V}_0 = 0.005 \cdot \mathbf{I}. \tag{14}$$

Согласно алгоритму фильтрации (12)÷(14) оптимальная оценка вектора состояния получается путем одновременного весового суммирования «обновляющих» процессов $\varepsilon_{n,t}$ всех каналов. Это свойство весового суммирования согласуется с интегральной способностью слуха воспринимать отклики различных слуховых каналов.

Структурная схема многоканальной слуховой фильтрации зашумленной речи, построенная в соответствии с соотношениями (12)÷(14) приведена на рис. 1.

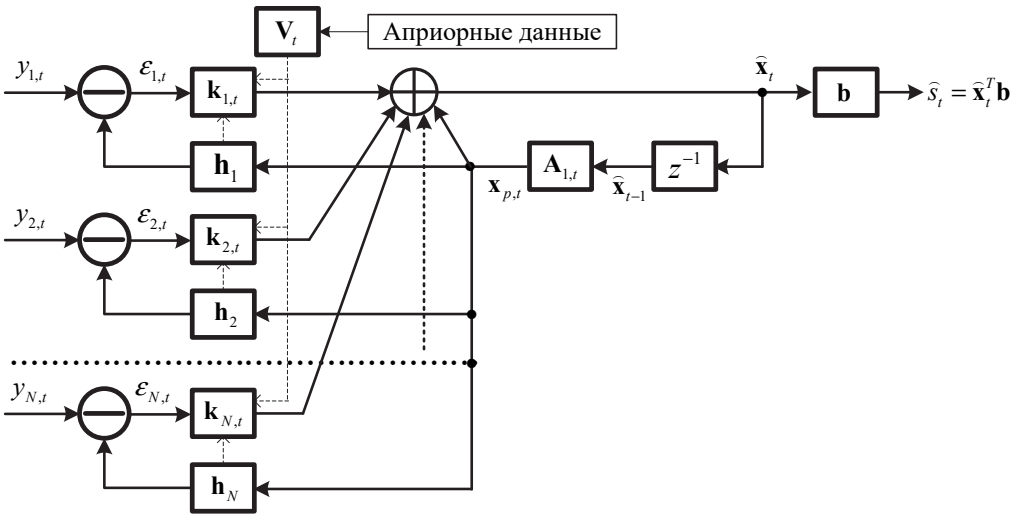

Рис. 1. Структурная схема многоканальной слуховой фильтрации зашумленной речи.

Принцип работы данной схемы не требует пояснений. Отметим лишь, что отклик системы является скалярной последовательностью оценок сигнала модели речеобразования в (1), определяемых по соотношению: $\widehat{s}_t = \widehat{\mathbf{x}}_t^T \mathbf{b}, t = 1,2,\ldots$.

Исследование метода фильтрации и разборчивость речи. Алгоритм фильтрации зашумленной речи реализован в среде MATLAB. В качестве исходной речи взята стандартная фраза: «*Эти жирные сазаны ушли под палубу*». Частота дискретиза-

ции 8 кГц. Число каналов 32. По средней дисперсии сигнала $D_s = \overline{D_{s,t}}$ и среднеквадратичной погрешности фильтрации $D_e = \overline{(s_t - \hat{s}_t)^2}$ проводился расчет ОСШ: $hdB = \overline{D}_s / D_e$, на основе которого определялся уровень $L(hdB)$ [3]:

$$L(hdB) = 0.0222 \cdot (hdB + 30) / \sqrt{1 + \left[(hdB + 30)/97.5\right]^6}\ . \qquad (15)$$

Слоговая разборчивость речи определялась по соотношению [3]:

$$S(\%) = 100\{1 + \nu\}\,\mathrm{x}\,\{1 - \nu\}, \quad \nu = 11{,}75635 \Big/ \left(1 + 10^{0.1 \cdot L(hdB)}\right), \qquad (16)$$

Результаты экспериментального исследования приведены на рис. 2.

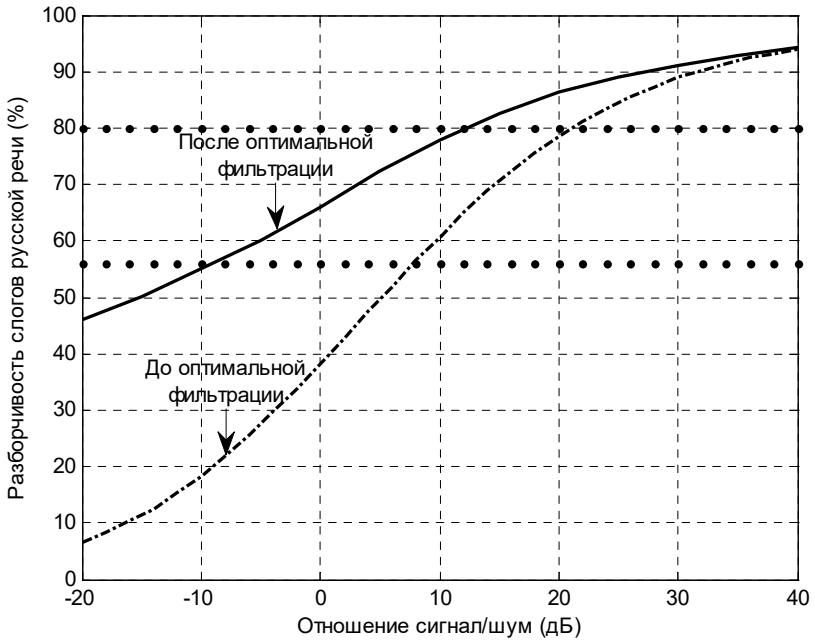

Рис. 2. Зависимости разборчивости речи от ОСШ.

Заключение. Рассмотренный метод многоканальной слуховой фильтрации и конструктивный алгоритм его реализации при отношениях сигнал/шум в диапазоне от минус 10 дБ и более согласно ГОСТ Р 51061-97 позволяет использовать этот метод фильтрации для работы в телефонных сетях общего пользования, обеспечивающих класс качества по слоговой разборчивости не ниже первого. Основная проблема данного метода фильтрации – сложность реализации. Разрешение этой проблемы - задача последующих исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Слуховая система*. / Ред. Я.А. Альтман. – Л.: Наука, 1990. – 620 с. – (Основы современной физиологии).
2. *Санников В.Г.* Теоретический анализ заметности искажений речевых сигналов по громкости их слухового восприятия // ЭЛЕКТРОСВЯЗЬ. – 2002. – № 12. – С. 38–42.
3. *Санников В.Г.* Метод оперативной оценки разборчивости речи по рабочим характеристикам слуховой системы // ТЕОРИЯ И ТЕХНИКА РАДИОСВЯЗИ, ОАО «Концерн «Созвездие». – 2009. – № 3. – С. 40-45.
4. *Леонов Ю.П.* Теория статистических решений и психофизика.–М.:Наука,1977.–223 с.
5. *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.
6. *Сейдж Э., Мелс ДЖ.* Теория оценивания и её применение в связи и управлении / Пер. с англ. // Под ред. Б.Р. Левина. – М.: Связь, 1976. – 495 с.