ФОРМИРОВАНИЕ БАНКА АПРИОРНЫХ ДАННЫХ О РЕЧИ ДИКТОРА

ВЛАДИМИР САННИКОВ ЮРИЙ ПРОХОРОВ ЮРИЙ ЖУРАВСКИЙ

Кафедра теории передачи сигналов, Московский электротехнический институт связи, Москва, СССР III024

АННОТАЦИЯ

Рассмотрена нелинейно-параметрическая модель речевого сигнала (РС). Дана методика оценки априорных данных, необходимых для оптимальной фильтрации РС. Приведены результаты исследования влияния априорных данных на качество фильтрации, а также зависимости их от длительности обучающей выборки, смыслового содержания и способа произнесения речи диктора.

МОДЕЛЬ СИГНАЛА И АПРИОРНЫЕ ДАННЫЕ

Современное состояние теории и техники цифровой обработки и передачи речи характеризуется широким использованием марковских моделей РС [1,2]. На основе концепции переменных состояния динамической системы, К которой отнесем систему речеобразования, запишем марковскую модель РС в виде

$$\begin{aligned} x_{t+1} &= F(a_t)\cdot x_t + G y_t + B_x \cdot w_{x,t}, \\ a_{t+1} &= \Lambda_a \cdot a_t + \lambda_o + B_a \cdot w_{a,t}, \\ y_{t+1} &= F(c_t)\cdot y_t + \Psi(y_t) + B_y \cdot w_{y,t} \\ c_{t+1} &= \Lambda_c \cdot c_t . \end{aligned} \qquad (1)$$

Первое уравнение отображает голосовой тракт, второе – управляющие процессы или сообщения, характеризующие изменение во времени параметров голосового тракта; третье – источник квазипериодического возбуждения тракта; четвертое – параметры источника квазипериодического возбуждения.

В (I) t – дискретное время, x_t – вектор отсчетов РС, a_t – вектор отсчетов сообщения, y_t – вектор отсчетов квазипериодического сигнала возбуждения, c_t – вектор параметров источника квазипериодического возбуждения, $F(a_t)$ – квадратная матрица с элементами $F_{i,i+1} = 1$, $i = \overline{1, m-1}$, $F_{m,i} = a_t^{(i)}$, $i = \overline{1, m}$, остальные её элементы равны 0, m – размерность векторов x_t и a_t; квадратная матрица $F(c_t)$ строится аналогично; Λ_a – квадратная матрица, характеризующая корреляционные взаимосвязи марковских сообщений; λ_o – постоянный вектор, который вместе с матрицей Λ_a характеризует среднее значение a_t; $\Psi(y_t)$ – нелинейная функция, обеспечивающая квазипериодический характер сигнала y_t [2]; G, B_x, B_a, B_y, Λ_c – постоянные матрицы; $w_{x,t}$, $w_{a,t}$, $w_{y,t}$, – случайные некоррелированные векторы гауссовских шумов с нулевыми средними и единичными матрицами корреляций. Для полного статистического описания поведения системы речеобразования необходимо задать начальные значения векторов средних значений и корреляционных матриц переменных состояния.

Объединяя переменные x_t, a_t, y_t, c_t в один блочный вектор состояния

$$z_t^T = (x_t^T; a_t^T; y_t^T; c_t^T)$$

запишем уравнение наблюдения в виде

$$u_t = h(z_t) + v_t, \qquad (2)$$

где u_t – наблюдаемая последовательность, $h(z_t)$ – скалярная функция векторного аргумента, v_t – случайная последовательность гауссовского шума наблюдений с нулевым средним и заданной дисперсией.

Соотношения (I) и (2) позволяют применить алгоритмы марковской фильтрации для выделения x_t из u_t с одновременным оцениванием процессов a_t, y_t, c_t. При этом синтез оптимальных алгоритмов фильтрации возможен только при известных характеристиках РС и его параметров. К таким априорным данным о речи диктора относятся матрицы: G, B_x, Λ_a, λ_o, B_a, B_y, Λ_c.

Отсутствие достаточно полных сведений о структуре и параметрах системы управления движением артикуляционного аппарата не позволяет воспользоваться моделью (I), так как указанные матрицы заранее не известны. Поэтому встает задача экспериментального их определения.

В качестве приближения к указанным априорным данным предлагается использовать их локально-постоянные оценки, которые можно получить при обработке фрагмента записи РС заданного источника в отсутствии помех. Совокупность отсчетов фрагмента РС образует обучающую выборку. Собственно процедура оценки элементов указанных выше матриц и называется формированием банка априорных данных о речи диктора.

МЕТОДИКА ОЦЕНКИ АПРИОРНЫХ ДАННЫХ

Фрагмент незашумленного РС разбивается на сегменты, каждый длительностью T_c = 10÷20 мс. Предполагается, что оцениваемые величины постоянны на сегменте анализа. Каждый сегмент обучающей выборки классифицируется по признаку "тон-шум". На участках типа "тон" оцениваются средние периоды основного тона T_0 и первой форманты T_1. Если c_t двумерный вектор, то его составляющие $c_t^{(1)}$ и $c_t^{(2)}$ можно выбрать так, чтобы при заданных вектор-функции $\Psi(y)$ и коэффициентах B_Ψ и B_y дискретное векторное колебание $\hat{y}_t$ имело собственную частоту $\hat{F}_1 = 1/T_1$ на периоде $\hat{T}_0 = T_0$. Далее по оценкам $\hat{c}_t$ оцениваются элементы λ_{cij} матрицы $\hat{\Lambda}_c$.

Оценка вектора коэффициентов $\hat{a}_t$ производится по обучающей выборке на основе известных алгоритмов идентификации параметров стохастических систем [3,4]. Траектории оценок $\hat{a}_t$ являются тем материа-

лом, который необходим для вычисления оценок $\hat{\Lambda}_a$, $\hat{\lambda}_o$, $\hat{B}_a$. С этой целью по траекториям $\hat{a}_t$ вычисляются векторы средних $\bar{\hat{a}}$ и стандартных отклонений $\hat{\sigma}_a$. Переходя к центрированным и нормированным траекториям $\tilde{a}_t^{(i)}=(\hat{a}_t^{(i)}-\bar{\hat{a}}^{(i)})/\hat{\sigma}_a^{(i)}$, $i=\overline{1,m}$, полагаем, что они удовлетворяют стохастическому уравнению

$$\tilde{a}_{t+1}=\tilde{\Lambda}_a\cdot\tilde{a}_t+\tilde{B}_a\cdot w_{a,t}. \qquad (3)$$

Полагая $\tilde{a}_t$ эргодическими процессами, определим усреднением по времени взаимнокорреляционные матрицы

$$E\tilde{a}_{t+1}\cdot\tilde{a}_{t+1}^T=\tilde{\Lambda}_a\cdot R_a^{(1)}+\tilde{B}_a\cdot\tilde{B}_a^T=R_a^{(0)},$$
$$E\tilde{a}_t\cdot\tilde{a}_{t+1}^T=\tilde{\Lambda}_a\cdot R_a^{(0)}=R_a^{(1)}, \qquad (4)$$

откуда находим

$$\tilde{\Lambda}_a=R_a^{(1)}\cdot[R_a^{(0)}]^{-1},$$
$$\tilde{B}_a\cdot\tilde{B}_a^T=R_a^{(0)}+R_a^{(1)}\cdot[R_a^{(0)}]^{-1}\cdot R_a^{(1)}. \qquad (5)$$

На основе (3) и (5) с учетом $\bar{\hat{a}}$ и $\hat{\sigma}_a$ получаем искомые оценки матриц в виде

$$\hat{\Lambda}_a=\mathcal{D}_\sigma\cdot\tilde{\Lambda}_a\cdot\mathcal{D}_\sigma^{-1},$$
$$\hat{\lambda}_o=\bar{\hat{a}}\cdot(1-\hat{\Lambda}_a);\quad \hat{B}_a=\mathcal{D}_\sigma\cdot\tilde{B}_a, \qquad (6)$$

где $\mathcal{D}_\sigma=diag[\hat{\sigma}_a^{(1)}\dots\hat{\sigma}_a^{(m)}]$ – диагональная матрица стандартных отклонений.

Важным вопросом методики оценки априорных данных является расчет необходимой длительности обучающей выборки. Она равна $T_{ов}=M\cdot T_c$, где M – минимально необходимое число сегментов на фрагменте анализа. Установлено, что $M\approx 15\cdot\Delta_g^{-2}$, где Δ_g – доверительный интервал для элементов матриц $R_a^{(o)}$ и $R_a^{(1)}$. При $\Delta_g=0{,}1$, $T_{ов}=10\div20$ с. Дальнейшее увеличение $T_{ов}$ не приводит к значительному изменению оценок данных.

АНАЛИЗ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ АПРИОРНЫХ ДАННЫХ

С целью выявления влияния априорных данных на качество фильтрации, а также анализа чувствительности их к способу речеобразования, были проведены различные эксперименты. Анализ результатов фильтрации зашумленного РС заданного диктора показал, что при отношении сигнал-шум 0 дб. алгоритмы фильтрации могут быть неработоспособными при пренебрежении априорными данными о его голосе. Алгоритм совместной фильтрации РС и оценки сообщения обеспечивает увеличение отношения сигнал-шум, но не очень значительно. Лучшие результаты получаются при совместном оценивании параметров, фильтрации и обнаружении РС. Такая обработка обеспечивает выигрыш в отношении сигнал-шум на 6–9 дб. при вероятности ошибки обнаружения 10^{-2}.

Коротко изложим результаты анализа чувствительности априорных данных к способу речеобразования, смысловому содержанию и индивидуальным особенностям голосов различных дикторов. Некоторые аспекты этой задачи рассмотрены в [5]. Эксперимент состоял в следующем. Подбирались фонограммы (обучающие выборки) разных дикторов. Для каждой фонограммы оценивались траектории сообщений $\hat{a}_{k,t}$, $k=\overline{1,N_\varphi}$, где N_φ – число фонограмм. Для установления факта однородности различных траекторий производилась проверка гипотезы о принадлежности всех траекторий генеральной совокупности. Если гипотеза справедлива, то различия в траекториях оцениваемых сообщений считаются незначительными и, следовательно, они однородны. В противном случае неоднородны. Это происходит тогда, когда априорные данные чувствительны к способу формирования фонограмм (темп речи, смысловое её содержание, индивидуальность голоса и др.).

Методика проверки истинности гипотезы состоит в следующем. На основе преобразования Фишера по случайным элементам r_{ij} корреляционных матриц $R_a^{(o)}$ и $R_a^{(1)}$ вычисляются новые случайные величины

$$z_{kij}=\frac{1}{2}\cdot\ln[(1+r_{kij})/(1-r_{kij})]. \qquad (7)$$

Определяются усредненные по всем фонограммам величины

$$\bar{z}_{ij}=\sum_{k=1}^{N_\varphi}\alpha_k\cdot z_{kij}\Big/\sum_{k=1}^{N_\varphi}\alpha_k, \qquad (8)$$

где $\alpha_k=M_k/3-3$. Затем формируется случайная величина взвешенного среднеквадратичного отклонения

$$\Delta z_{ij}^2=\sum_{k=1}^{N_\varphi}\alpha_k(z_{kij}-\bar{z}_{ij})^2, \qquad (9)$$

имеющая хи-квадрат распределение с $N_\varphi-1$ степенями свободы. Теперь, если $P\{\chi^2>\Delta z_{ij}^2\}>\alpha$, то принимается гипотеза об однородности различных фонограмм. Здесь α – уровень значимости. Эксперимент проводился на речевом материале удовлетворяющем ГОСТ 16600-72. Параметры обработки РС: частота дискретизации 15 кГц, $T_c=15$ мс, $m=2$ и 4.

Анализ экспериментальных данных показал, что априорные данные, полученные для разных дикторов на одной и той же фразе в естественных условиях произнесения с $\alpha=0{,}001$ надежно различаются. Способ произнесения и смысловое содержание фонограммы практически не влияют на априорные данные.

В заключении отметим, что проведенный анализ эффективности использования априорных данных подтверждает практическую важность решения поставленной задачи. При этом априорные данные целесообразно формировать для каждого конкретного диктора или группы дикторов со сходными голосами.

ЛИТЕРАТУРА

[1] Ю.Н. Прохоров, "Статистические модели и рекуррентное предсказание речевых сигналов" – Радио и связь: М., 1984.

[2] М.В. Назаров, Ю.Н. Прохоров, "Методы цифровой обработки и передачи речевых сигналов" – Радио и связь: М., 1985.

[3] R. Lee. "Optimal Estimation, Identification and Control." Gamburg, Massachusetts, 1966.

[4] В.Г. Санников, Ю.И. Журавский, Ю.Н. Прохоров, "Формирование банка данных о речи диктора", Тезисы докл. Всесоюзного семинара АРСО-12, Киев-Одесса, 1982.

[5] М.В. Назаров, Ю.Н. Прохоров, Ю.И. Журавский, "Исследование характеристик параметров авторегрессии заданных источников", Тезисы докл. Всесоюзного семинара АРСО-13 Новосибирск, 1984.