УДК 621.391

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

АЛГОРИТМЫ РАСПОЗНАВАНИЯ ФОНЕМ РЕЧИ

Минимальной речевой единицей является фонема – абстрактное обозначение конкретного звука устной речи. Поскольку некоторые звуки существенно нестационарные, то говорят о середине фонемы – точке, в которой спектр наиболее стабилен. Трудности фонемного анализа, связанные с коартикуляцией или взаимным влиянием соседних фонем, иногда вызывают необходимость сегментации речи на более крупные элементы – дифоны (диады) и трифоны (триады). Дифоном называют участок речи, расположенный между центрами двух соседних фонем. Трифон – конечная половина первой фонемы, вторая фонема и первая половина третьей фонемы.

Главным преимуществом распознавания речи на основе речевых (смысловых) единиц является ограниченное число базовых элементов. Так, число фонем для различных языков – 30–50, минимальное число дифонов 600–1200, а трифонов около 3000.

Одной из важных задач обработки речи является распознавание отдельных фонем речи различных дикторов. Трудности построения систем автоматического распознавания речи, работающих на множестве дикторов (дикторонезависимых), связаны с органической разницей параметров голосового тракта и источником возбуждения у различных дикторов при произнесении одних и тех же звуков речи. Это приводит к различию частоты основного тона, формантных частот и других параметров и изменению этих параметров во времени. Параметры человеческого голоса, приводящие к изменению характеристик речевой волны, меняются в зависимости от возраста, а также эмоционального и психологического состояния человека (усталости, болезни и т.д.).

Синтез алгоритма выполнен по векторному критерию при учете совокупности показателей качества распознавания фонем: вероятности правильного распознавания и устойчивости распознавания.

Один из подходов к решению задачи распознавания речи основан на сопоставлении оценок параметров речевых структурных единиц, в виде непрерывной во времени выборки либо последовательности выборок. Существующие методы решения задач распознавания речи основаны на изучении структуры речи и её анализе с использованием математического аппарата различного вида и уровня сложности.

Для сегментации используются разные критерии, 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, не известно достаточно эффективное решение данной проблемы.

Целью данной работы является синтез алгоритмов фонемного распознавания речи, устойчивых в условиях работы с разными дикторами.

Полагается, что на вход системы распознавания поступает временная последовательность отсчетов речевого сигнала, взятых с заданным интервалом дискретизации.

Обучающие выборки речи для каждого из дикторов заданы в виде классифицированных обучающих выборок. Считается, что время произношения фонем в слитном речевом сигнале априори неизвестно. Априорные вероятности появления конкретной фонемы для всей совокупности фонем одинаковы.

Качество алгоритма распознавания оценивается совокупностью показателей эффективности распознавания сигналов и устойчивостью алгоритмов к воздействию аддитивной помехи.

В качестве показателя эффективности используется средняя вероятность правильного распознавания фонем при низком уровне аддитивной помехи.

Под показателем устойчивости алгоритмов понимается значение средней вероятности правильного распознавания фонем $P_{пр}$ при воздействии аддитивной помехи в канале связи с

заданным отношением сигнал-шум q. Необходимо построить оптимальный алгоритм, который по предъявленной реализации речи выносит решения о принадлежности произнесенных слов неизвестных дикторов к заданным классам и обеспечивал бы максимум в классе робастных алгоритмов.

Поставленная задача распознавания включает следующие этапы – предварительную обработку, оценивание параметров обстановки, сегментацию речи, оценивание признаков, отбор и хранение эталонов, а также принимается решение об определенной речевой единице. Перед сегментацией проводится нормировка всего сигнала по энергии сигнала, которая позволяет снизить чувствительность алгоритмов сегментации к громкости речи, типу микрофона и ослабить влияние различий в количестве уровней квантования, акустоэлектрических настроек устройств предварительной обработки сигнала и т.д.

Решение задачи распознавания речи и ряда других задач во многом связано с успешным проведением сегментации речи на речевые единицы. Фонемная сегментация выполняется на основе модели речеобразования. Возможно последовательное или одновременное использование ряда признаков. При последовательном вначале производится деление звуковых сигналов на речь и паузу, далее на вокализованные и невокализованные, взрывные, глухие, звонкие и т.д. На последнем этапе производится фонемная сегментация. Некоторые алгоритмы сегментации описаны и исследованы в [1].

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[1].

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

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

При априорном знании ковариационных матриц R_n, R_{n-1} процедура принятия решения состоит в сравнении логарифма отношения правдоподобия E_n с порогом $k = \ln(c)$ априорно выбранного критерия качества.

Гипотеза о начале и конце нового сегмента фонем с ковариационной матрицей R_n принимается, если выполняется неравенство

$$E^r = \ln W(\hat{K}^r / R_{n-r1}) - \ln W(\hat{K}^r / R_{n+r1}) =$$

$$= 0{,}5sp(\hat{K}^r (R_{n+r1}{}^{-1} - R_{n-r1}{}^{-1})) \geq ln\, c - (\tau/2)\ln(\det R_{n+r1}/\det R_{n-r1}).$$

Поэтому алгоритм сегментации по множеству оценок ковариационных матриц сводится к виду

$$\sum_{l=0}^{p-1}\sum_{i=0}^{p-1} h(j,i)\hat{K}^r(i,j) \geq \Lambda_1, \text{ где } h(j,i) = R_{n+r1}^{-1}(j,i) - R_{n-r1}^{-1}(j,i),$$

где $R_n^{-1}(j,i), R_{n-1}^{-1}(j,i)$- корреляционные функции в соседних сегментах; р-порядок модели.

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

$$\sum_{i=0}^{N-1} S^k(i)H(i) < \Lambda_1 \text{ или } \sum_{i=0}^{N-1} S^k(i)H(i) \geq \Lambda_2.$$

Полагая оценки $\hat{K}^r(i,j) = R_n(j,i)$ и $\Lambda_1 = 0$, r1=r2=r, получим, что границам сегментов соответствуют локальные минимумы и максимумы функционала

$$D_1(f_n, f_{n-1}) = \ln(R_{n-r}R_{n+r}^{-1} / R_{n-r}R_{n-r}^{-1}).$$

Для задачи речевой обработки используются коэффициенты линейного предсказания (КЛП), определяемые по автокорреляционной последовательности фрагмента речевого

сигнала с помощью эффективной вычислительной процедуры Дарбина. При этом КЛП определяют с помощью алгоритма Дарбина из решения системы уравнений:

$$\sum_{j=0}^{p} r_n(i-j)a_n(j) = 0,$$

где p – порядок модели линейного предсказания; $r_n(i)$ – автокорреляционные коэффициенты на текущем n-м сегменте.

Алгоритм Левинсона позволяет определить и АР-параметры и коэффициенты отражения по заданной автокорреляционной последовательности. При этом АР-параметры порядка p для m-го блока вычисляют через АР-параметры порядка p как

$$a_p{}^{(m)}(n) = a_{p-1}{}^{(m)}(n) + k_p{}^{(m)} a_{p-1}{}^{(m)}(p-n).$$

Коэффициент отражения имеет вид

$$k_p{}^{(m)} = a_p{}^{(m)}(n) = -\sum_{n=0}^{p-1} a_{p-1}{}^{(m)}(n) r_{xx}{}^{(m)}(p-n) / \rho_{p-1}{}^{(m)}.$$

Дисперсия возбуждающего белого шума

$$\rho_p{}^{(m)}(n) = \rho_{p-1}{}^{(m)}(n)(1-|k_p{}^{(m)}|^2).$$

Вычисляют функционал

$$w^{(m)} = \ln(\sum_{n=1}^{p}(\rho^{(m)}(n) - \rho^{(m-1)}(n))^2).$$

Границы фонем определяют как временное положение локальных экстремумов при условии $w^{(m)} > P$, где P – порог.

Результат согласованной фильтрации можно сформировать, пропуская речевой сигнал через фильтр, коэффициентами которого являются параметры предсказания, либо через решетчатый фильтр, весовыми коэффициентами которого будут коэффициенты отражения.

Основное уравнение, описывающее структуру решетчатого фильтра с прямой связью (модели)

$$f_j(t) = f_{j+1}(t) - \rho(j) b_{j+1}(t-1), \qquad b_j(t) = b_{j+1}(t-1) - \rho(j) f_{j+1}(t).$$

Основное уравнение, описывающее структуру решетчатого фильтра с обратной связью

$$f_{j+1}(t) = f_j(t) - \rho(j+1) b_j(t-1), \qquad b_{j+1}(t) = b_j(t-1) - \rho(j+1) f_j(t).$$

Если порождающий сигнал модели приложен к решетчатому фильтру с прямой связью, а наблюдаемый речевой сигнал – решетчатому фильтру с обратной связью, то исходный порождающий сигнал восстанавливается.

Наблюдаемый речевой сигнал и коэффициентами отражения для фонемной сегментации берутся с разных соседних блоков, при этом для фонемной сегментации достаточно измерить дисперсию отклика фильтра.

Форма записи меры Итакура через параметры модели линейного предсказания при использовании автокорреляционного метода

$$D_1(f_n, f_{n-1}) = \ln(a_{n-r2} R_{n+\kappa 1} a_{n-r2}^T / a_{n+r1} R_{n-k2} a_{n+r1}^T),$$

где R – автокорреляционные коэффициенты на текущем n-м сегменте; a – КЛП, которые определяют с помощью алгоритма Дарбина или Левинсона[1-5].

Для небольших изменений спектра (этому условию можно всегда удовлетворить путем выбора соответствующей длины фрагмента речевого сигнала) вместо взятия логарифма в (2) используем следующее приближение:

$$D_M(f_n, f_{n-1}) = a_{n-r2} R_{n+\kappa 1} a_{n-r2}^T / a_{n+r1} R_{n-k2} a_{n+r1}^T - 1.$$

Квадратичная форма $a_{n-1}R_n\,a_{n-1}$ представляет собой минимальное значение среднеквадратической погрешности предсказания для текущего n-го сегмента.

Выражение можно представить в виде

$$D_M(f_n, f_{n-1}) = \sum_{m=0}^{p} V^{(n-r2)}(m)R^{(n+k1)}(m) / \sum_{m=0}^{p} V^{(n+r1)}(m)R^{(n-k2)}(m) - 1,$$

где $R(0) = \sum_{l=0}^{p}(a_l)^2$; $R(m) = 2\sum_{l=0}^{p-m} a_l \cdot a_{l+m}$; $l \le m \le p$.

Аналогично решение о изменении свойств речевого сигнала (начале нового и конце текущего сегмента фонем в очередной выборке) принимается по результату сравнения с порогом значений $d(f_n, f_{n-1})$, если

$$d(f_n, f_{n-1}) < \Lambda_1 \text{ или } d(f_n, f_{n-1}) > \Lambda_2,$$

где $d(f_n, f_{n-1}) = \log(\sum_{m=0}^{p} V^{(n-r2)}(m)R^{(n+k1)}(m) / \sum_{m=0}^{p} V^{(n+r1)}(m)R^{(n-k2)}(m))$.

При этом временные границы определяются по экстремумам предложенных функционалов.

Решение о начале нового и конце текущего сегмента фонем в очередной выборке принимается по результату сравнения с порогом значений $R_n^{фон}$ по формуле

$$R_n^{фон} < \Lambda_1 \text{ или } R_n^{фон} > \Lambda_2,$$

где $R_n^{фон} = \sum_{i=1}^{L(n)} \min_{j\in[-J,J]} \alpha_{i,j}^{l} \,|\, \hat{f}_i(n) - \hat{f}_{i+j}(n+1)|^q$ – функционалы, построенные на основе метрик в пространстве L1, L2; $\hat{f}_i(n)$ – 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.

Для сегментации возможно использование авторегрессионных формантных оценок. Авторегрессионая спектральная оценка формантных частот вычисляются в соответствии с выражением

$$\vec{f}_\nu = \frac{F_\partial}{N} \arg loc \max\{|1 + \sum_{n=1}^{p} \hat{a}(n)\exp(-j2\pi nk)|^{-1}, k = \overline{0, M}\}, \qquad (1)$$

где $\vec{f} = \arg loc \max(\vec{x})$ – векторная функция, задающая соответствие элементам входной последовательности $x_1, x_2, \ldots, x_N$ элементам выходной последовательности упорядоченное множество номеров локальных максимумов $\{ f_i \ , \ i = \overline{0, L} \}$; вектор оценок $\vec{f}_\nu = \{ f_{i,\nu}, i = \overline{0, L} \}$, L – количество локальных максимумов в спектре; $F_\partial = 1/\Delta t$ – частота дискретизации сигнала, Δt – период дискретизации сигнала; $M = Z[N/2-1]$; Z[y] – функция округления к ближайшему целому числу.

Псевдоформантные (модифицированных АР) оценки формантных частот вычисляются в соответствии с выражением

$$\vec{f}_\nu = \frac{F_\partial}{N} \arg loc \max\{|1 + \sum_{n=1}^{p-1} \hat{a}[n]\exp(-j2\pi nk) + \alpha\exp(-j2\pi pk)|^{-1}, k = \overline{0, M}\}, \qquad (2)$$

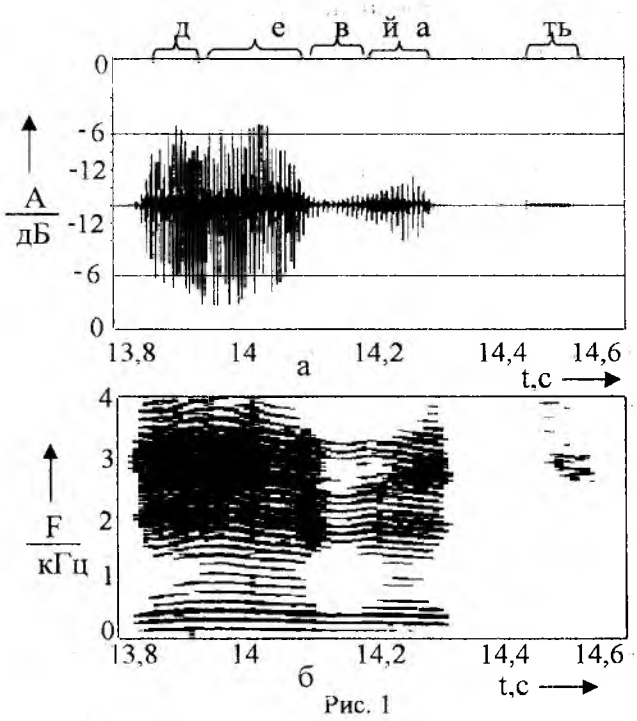

Рис. 1

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

Для оценивания формант по спектрально полосным признакам вычисляют спектрально-полосные сигналы, соответствующие вероятному расположению формант. Оценки формантных частот вычисляются как среднеэффективные частоты в заданных полосах частот

$$\hat{f}_э^{(m)} = \sum_{i=f_н^{(m)}}^{f_в^{(m)}} i(S_{э,i}^2) \Big/ \sum_{i=f_н^{(m)}}^{f_в^{(m)}} (S_{э,i}^2).$$

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

Предлагаемые алгоритмы определения были исследованы на словах речи. Ввод речевого файла осуществлялся через специализированный адаптер. Тактовая частота 8 Кгц, 16 разрядов на один отсчет. На интервалах речевого файла длительностью 15 мкс, взвешенного

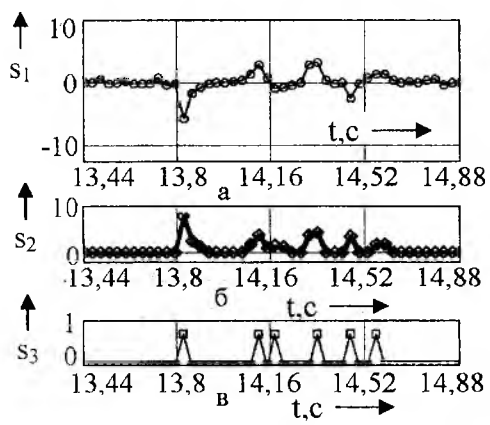

Рис. 2

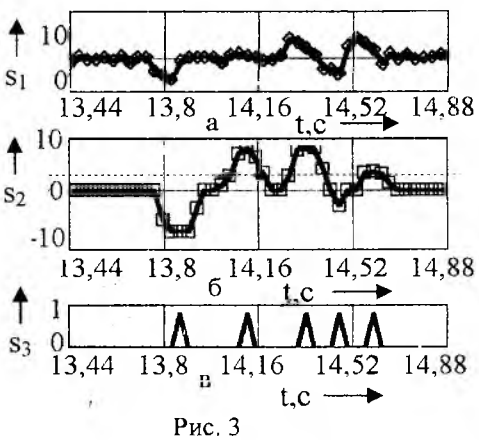

Рис. 3

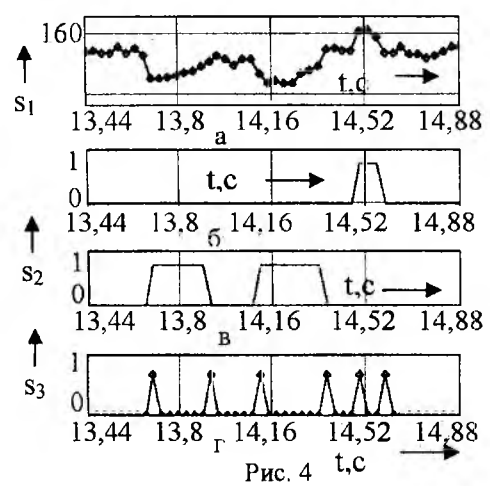

Рис. 4

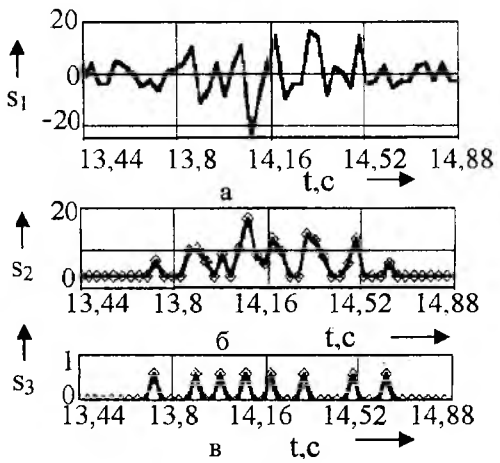

Рис. 5

окном Хэмминга 45 мкс, вычислялись 12 коэффициентов линейного предсказания (рис. 1а, 1б – огибающая и сонограмма слова «девять»; рис. 2а – статистики меры Уишарта для слова «девять»; рис. 2в – границы фонем, полученные из статистики меры Уишарта; рис. 3а – статистики меры Итакура для слова «девять»; рис. 3в – границы фонем, полученные из статистики меры Итакура; рис. 4а – статистика нуль пересечений для слова «девять; рис. 4б – невокализованность; рис. 4в – вокализованность; рис. 4г – границы вокализованных и невокализованных фонем [1]; рис. 5а – статистика $R_n^{фон}$ среднеэффективных частот в формантных полосах частот (двухэтапная) для слова «девять»; рис. 5б – статистика модуля среднеэффективных частот в формантных полосах частот (двухэтапная) для слова «девять» с учетом ограничения по порогу; рис. 5в – границы фонем, вычисленные для двухэтапного спектрально-полосного алгоритма), рис. 6а – статистика вычисленная через коэффициенты отражения с учетом ограничения по нижнему порогу для слова «девять»; рис. 6б – границы фонем, вычисленные для алгоритма вычисления границ через коэффициенты отражения).

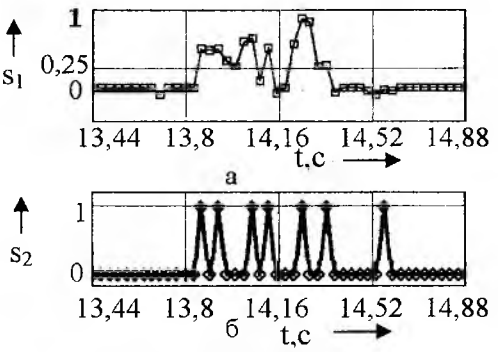

Рис. 6

После выполнения сегментации фонем необходимо принять решение о наибольшей степени близости в пространстве признаков произносимой фонемы и одной из фонем обучающих выборок. В качестве информативных параметров, используемых для распознавания, могут быть различные характеристики речевых сигналов. К ним относятся частота основного тона, формантные частоты[1, 2], признак вокализованности [1, 2], мощность сигнала в разных полосах частот сигнала[1, 2], длительности произносимых фонем. Наилучшим образом задача распознавания фонем может быть решена с использованием алгоритмов оценивания формантных признаков, которые характеризуют голосовой тракт. С целью получения динамических признаков распознаваемого цифрового сигнала производится разбиение слов на блоки одинаковой длительности, которое обычно составляет 10-30 мс. Выполненные нами исследования показали, что оценки формантных признаков существенно изменяются при сдвиге всего сегмента лишь на один дискрет. Это требует особой точности сегментации либо поиска алгоритмов, устойчивых к подобным ошибкам. Существенно понизить чувствительность к временному сдвигу удается путем предварительной нерекурсивной цифровой ВЧ коррекции [1] либо использованием алгоритмов выбеливания речевых сигналов.

Для оценивания значений формантных частот рационально использовать двухэтапные спектрально-полосные алгоритмы, двухэтапные алгоритмы с определением количества нуль-пересечений в выделенных полосах частот, алгоритмы кепстрального оценивания, классические методы спектрального оценивания и алгоритмы линейного предсказания речевых сигналов [1, 2].

Подход к авторегрессионому оцениванию формантных частот строится на основе модели линейного предсказания речевых сигналов. Соответствующий алгоритм предполагает следующую последовательность шагов.

Вначале оценивается корреляционная функция и методом Левинсона вычисляются оценки коэффициентов авторегрессии. Корреляционная функция полученных эталонов может пересчитываться с учетом статистических свойств шума на этапе распознавания. В случае независимости сигнала и помехи корреляционная матрица, смеси $R = R_э + R_п$, где $R_э$ – исходная корреляционная матрица; $R_п$ – корреляционная матрица помехи. При действии белого шума высокого уровня на входе устройства скорректированная корреляционная матрица эталона $R = R_э + I(\sigma_c^2 - \sigma_э^2)$, где $R_э$ – исходная корреляционная матрица;

I – единичная матрица, $\sigma_c^2, \sigma_э^2$ – дисперсия сигнала на входе устройства распознавания и эталона соответственно. Затем определяется авторегрессионая спектральная оценка формантных частот в соответствии с выражениями (1) или (2). Псевдоформантные оценки (2) обладают более выраженными (подчеркнутыми) формантами из-за проектирования корней характеристического уравнения на единичный круг комплексной плоскости.

Согласно алгоритму оценивания формант по спектрально-полосным признакам вычисляют спектрально-полосные сигналы, соответствующие вероятному расположению формант. Граничные частоты $f_в^{(m)}$, $f_н^{(m)}$ полосы, соответствующие m-м формантам при частоте дискретизации 8 кГц, приведены в табл. 1.

Адаптивные оценки формантных частот с учетом разной помеховой обстановки на этапе обучения и распознавания строятся следующим образом. Перед распознаванием и обучением проводится нормировка всего сигнала по мощности помехи в паузе. На этапе обучения и распознавания вычисляются формантные частоты как среднеэффективные частоты в заданных полосах частот

$$\hat{f}_э^{(m)} = \sum_{i=f_н^{(m)}}^{f_в^{(m)}} i(S_{э,i}^2 + A(q_{тек}))) / \sum_{i=f_н^{(m)}}^{f_в^{(m)}} (S_{э,i}^2 + A(q_{тек})), \quad (4)$$

на этапе распознавания

$$\hat{f}_{тек}^{(m)} = \sum_{i=f_н^{(m)}}^{f_в^{(m)}} i(S_i^2 + G_i C(1 - \mathrm{sgn}(q_{тек} - 1))) / \sum_{i=f_н^{(m)}}^{f_в^{(m)}} (S_i^2 + G_i C(1 - \mathrm{sgn}(q_{тек} - 1))), \quad (5)$$

где ($f_в^{(m)}$, $f_н^{(m)}$) – диапазон частот для m-й форманты, S_2 – вычисленный дискретный спектр; $q = (\sum_{i=\tau_{c,1}}^{\tau_{c,2}} (S_i)^2 / (\tau_{c,2} - \tau_{c,1} + 1)) / (\sum_{i=\tau_{n,1}}^{\tau_{n,2}} (S_i)^2 / (\tau_{n,2} - \tau_{n,1} + 1))$ – оценка отношения сигнал-шум на этапе обучения и распознавания соответственно; $A(q_{тек}) = G_i(B / q_{тек}^2 + C(1 - \mathrm{sgn}(q_{тек} - 1))$ – корректирующая функция; G_i – весовая функция с учетом спектра помехи и характеристики фильтра предварительной обработки речевого сигнала; функция $\mathrm{sgn}(u) = \begin{cases} 1, если\ u \geq 0, \\ 0, если\ u < 0. \end{cases}$

Таблица 1

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

Аналогично оценки формантных частот могут вычисляться путем подсчета количества нуль 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 двукратным либо трехкратным перекрытием или без него.

Оценки формантных частот спектрально-полосным методом могут вычисляться с использованием АР оценок 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. Экспериментальные исследования показали, что наибольшие вероятности правильного распознавания структурных речевых единиц получаются для авторегрессионых спектральных оценок частот в каждой из 4-х формантных полос при порядке модели p=3.

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 $\sum_{r=-v}^{u} W_r = 1$.

Процедура вычисления формант может быть повторена, но при этом в качестве граничных полос частот используют

$$\hat{f}_{в}^{(m)} = \hat{f}^{(m)} + \Delta, \quad \hat{f}_{н}^{(m)} = \hat{f}^{(m)} - \Delta, \tag{6}$$

где $\hat{f}^{(m)}$ – форманты вычисленные на предыдущем этапе, Δ - границы диапазона поиска формант. Простейшей среди рекуррентных процедур является двухэтапная.

Методы вычисления формантных частот основаны на параметрическом либо непараметрическом оценивании спектра и отборе либо усреднении локальных максимумов спектра.

Обычно весь набор эталонов для одной из фонем разделяют на классы, и каждый новый диктор, обращающийся к системе, пользуется эталонами того диктора (участвовавшего в обучении) или класса дикторов, параметры голоса которого наиболее близки его собственным параметрам.

Решение о фонеме на основе модели в виде смеси распределений авторегрессионых и спектрально-полосных оценок формантных частот принимается из условия

$$i = \arg\max_{l=\overline{0,M}} \Big(\sum_{d=1}^{D} \sum_{s=0}^{S(d)} (k_{KosAR} \cdot \ln(u_{пр} R_{пр}^{(l,s,d)} + u_{обр} R_{обр}^{(l,s,d)}) + k_{KosSP} \cdot \ln(R^{(l,s,d)})) \Big), \tag{7}$$

где $R_{пр}^{(l,s,d)}, R_{обр}^{(l,s,d)}$ – функционалы авторегрессионых оценок, использующие считывание данных от начала и конца фонемы, $R^{(l,s,d)}$ – функционал, построенный на основе спектрально-полосных оценок.

Для полигауссовского распределения формантных частот статистики $R_{пр}^{(l,s,d)}$ вычисляют по оценкам формантных частот $\hat{f}_{i,пр}(n)$ и $\hat{f}_{i+j,пр}^{l,s}(n)$, а $R_{обр}^{(l,s,d)}$ – по оценкам $\hat{f}_{i,обр}(n)$ и $\hat{f}_{i+j,обр}^{l,s}(n)$ в виде

$$R^{(l,s,d)} = \sum_{d=1}^{D} \sum_{s=0}^{S(d)} P_{в}(\omega_{s,d}) \prod_{n=1}^{N} \prod_{i=1}^{L(n)} P_{i,n}^{(l,s,d)}, \tag{8}$$

где $P_{i,n}^{(l,s,d)} = \sum_{j=-J}^{J} \sum_{h=-H}^{H} P_{c}(\omega_{j,h}) (2\pi)^{-0,5} |\sigma_s|^{-1} \exp(-|\hat{f}_{i}(n) - \hat{f}_{i+j}^{l,s}(n+h)|^2 / (2\sigma_s^2))$.

При равновероятном законе $P_{в}(\omega_{s,d})$, $P_{c}(\omega_{j,h})$ можно положить, что вероятности $P_{в}(\omega_{s,d}) = 1/((2J+1)(2H+1))$, $P_{c}(\omega_{j,h}) = (\sum_{d=1}^{D} S(d))^{-1}$.

Для модели в виде смеси распределений Коши статистики $R_{пр}^{(l,s,d)}$ вычисляют по оценкам формантных циклических частот $\hat{\omega}_{i,пр}(n)$ и $\hat{\omega}_{i+j,пр}^{l,s}(n)$, а $R_{обр}^{(l,s,d)}$ по оценкам $\hat{\omega}_{i,обр}(n)$ и $\hat{\omega}_{i+j,обр}^{l,s}(n)$ в виде

$$R^{(l,s,d)} = \sum_{d=1}^{D} \sum_{s=0}^{S(d)} P_в(\omega_{s,d}) \prod_{n=1}^{N} \prod_{i=1}^{L(n)} \sum_{j=-J}^{J} \sum_{h=-H}^{H} P_c(\omega_{j,h}) \frac{c_i(n)}{\pi(c_i(n)^2 + |\hat{\omega}_i(n) - \hat{\omega}_{i+j}^{l,s}(n+h)|^2)}, \quad (9)$$

где $c_i(n) = \Delta\omega_i(n)/2$ – параметр распределения Коши, $\Delta\omega$ – циклическая полоса частот формант.

Таблица 2

Алгоритм	$\hat{P}_{п.ср.}$
СП2 и АР Коши	0,925
АР Коши	0,72
СП2 Коши	0,9
СП2 Полигаусс	0,8

Ширина полос формант изменяется в зависимости от произносимых фонем в пределах от 50 до 400 Гц. В некоторых случаях можно положить, что полосы формант одинаковы, тогда $c_{i,np}(n) = \Delta\omega/2$.

На основе синтезированных алгоритмов создан многофункциональный исследовательский пакет распознавания слов, с применением которого выполнены статистические испытания алгоритмов распознавания слов, веденных в ЭВМ с микрофона через звуковой интерфейс с частотой дискретизации 8 кГц.

В экспериментальных исследованиях получено, что сочетание алгоритмов, например авторегрессионых и спектрально-полосных на основе смесей Коши, даже для одного эталона дает выигрыш в вероятности правильного принятия решения (табл. 2), где СП2 – двухэтапные спектрально-полосные, а АР – авторегрессионые оценки формант. При этом в эксперименте полагалось, что коэффициенты $k_{KosAR} = 1, k_{KosSP} = 1$. Вероятности перепутывания

Таблица 3

р	и	ы	о	ю	я	е	ё	у	а	э
и	0,80	0,13	0	0	0	0	0	0	0	0
ы	0,16	0,87	0	0	0	0	0	0	0	0
о	0	0	0,98	0	0	0	0	0	0	0
ю	0	0	0	0,96	0	0	0,07	0	0	0
я	0	0	0	0	1	0	0	0	0	0
е	0,04	0	0	0	0	0,98	0	0	0	0
ё	0	0	0	0,04	0	0	0,93	0	0	0
у	0	0	0	0	0	0	0	1	0	0
а	0	0	0	0	0	0,02	0	0	1	0
э	0	0	0,02	0	0	0	0	0	0	1

Таблица 4

р	и	ы	о	ю	я	е	ё	у	а	э
и	0,76	0,13	0	0	0	0	0	0	0	0
ы	0,16	0,87	0	0	0	0,02	0	0	0	0
о	0	0	0,98	0	0	0	0,07	0,07	0	0
ю	0	0	0	1	0	0	0,13	0	0	0
я	0	0	0	0	1	0	0	0	0	0
е	0,08	0	0	0	0	0,98	0	0	0	0,02
ё	0	0	0	0	0	0	0,80	0	0	0
у	0	0	0	0	0	0	0	0,93	0	0
а	0	0	0	0	0	0	0	0	1	0
э	0	0	0,02	0	0	0	0	0	0	0,98

гласных букв для алгоритмов распознавания построенных на основе совместного использования спектрально-полосных и авторегрессионых смесей Коши показаны в табл. 3, а спектрально-полосных смесей Коши в табл. 4.

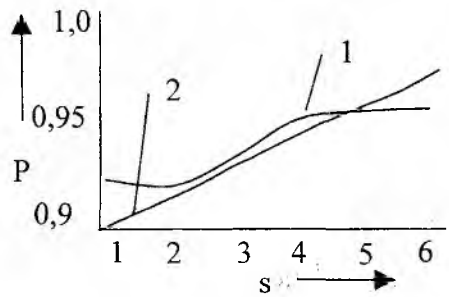

Рис. 7

По найденным рабочим характеристикам проведены сравнительные исследования алгоритмов распознавания. Экспериментально полученные результаты подтвердили высокое качество дикторонезависимых адаптивных алгоритмов распознавания по одному эталону на основе совместного использования спектрально-полосных и авторегрессионых смесей Коши при распознавании разных дикторов со средней вероятностью правильного распознавания 0,925 (табл. 2).

На рис. 7 показана зависимость вероятности правильного распознавания P от количества эталонов s для алгоритмов на основе смесей Коши: 1 – совместного использования авторегрессионых и спектрально-полосных; 2 – спектрально-полосных. При этом для случая использования

спектрально-полосных смесей Коши вероятность правильного распознавания при 6 эталонах 0,967.

На рис. 8 показана зависимость вероятности правильного распознавания от отношения сигнал-шум q спектрально-полосного алгоритма в виде смеси распределений Коши при априори известной сегментации речи, адаптивном решении задачи и действии шума на этапе принятия решений.

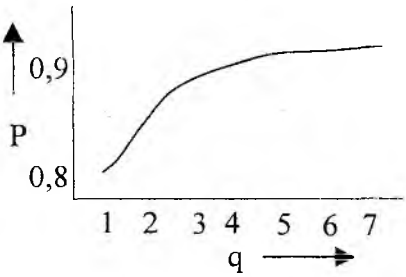

Рис. 8

Таким образом, в настоящей работе разработаны алгоритмы дикторонезависимых адаптивных алгоритмов распознавания фонем речи для полигауссовых распределений и смесей Коши. Результаты исследования предложенных алгоритмов распознавания речи на основе полигауссовых распределений и смесей Коши показали, что они характеризуются более высоким качеством распознавания, чем по одному эталону.

Список литературы: 1. *Пресняков И.Н., Омельченко С.В.* Помехоустойчивые алгоритмы сегментации речи в системах обработки // Радиотехника: Всекур. межвед. науч.-техн. сб. 2003. Вып. 131. С. 165 – 177. 2. *Пресняков И.Н., Омельченко А.В., Омельченко С.В.* Автоматическое распознавание речи в каналах передачи // Радиоэлектроника и информатика: науч.-техн. журн. 2002. №1. С. 26 – 31. 3. *Дж. Д. Маркел, А. Х. Грей.* Линейное предсказание речи. М.: Связь, 1980. 308 с. 4.*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. 5. *Марпл. – Мл. С. Л.* Цифровой спектральный анализ и его приложения. М.: Мир, 1990. 584 с.

Харьковский национальный
университет радиоэлектроники

Поступила в редколлегию 22.07.2003