УДК 681.142.66

Ниценко А.В., Шелепов В.Ю.

Институт проблем искусственного интеллекта

Украина, 340048, г. Донецк, ул. Артема, 118-б

E-mail: shel@iai.donetsk.ua

Алгоритмы пофонемного распознавании слов наперед заданного словаря

В статье описывается система пофонемного распознавания. Она опирается на новую методику сегментации, развивающую подходы, предложенные в работах [1], [2]. Новым моментами являются: использование при сегментации лишь обобщенной транскрипции, описывающей чередование гласных и голосовых согласных звуков, отказ от использования для разделения каждой пары фонем своего фильтра и использование таблиц типа той, которая представлена на рисунке 4. Это, естественно, повышает дикторонезависимость сегментации. Предложены новые подходы для выделения в сигнале звука “р” и для разделения нескольких идущих подряд согласных звуков. Все приведенные рисунки, иллюстрирующие работу описываемых алгоритмов, являются образами окон реальной программы распознавания, использующей эти алгоритмы.

The article is devoted description of phoneme speech recognition system. It based on new segmentation methods, which develop [1], [2] approaches. New moments: using only generalized transcription which reflects vowel and consonant interchange, rejection of using special filter for every pair of phonemes and using tables of picture 4 type. It raises speaker independence. New detection of “r” is proposed. All illustrations are images of real recognition-program windows.

1. Общая схема работы распознавателя

Речевой сигнал, поступающий с микрофона, оцифровывается как 8-битный, с частотой дискретизации 22050 Гц. Мы будем на всех этапах обработки разбивать оцифрованный сигнал на непересекающиеся отрезки длиной 256 отсчетов. Ниже слово «отрезок» понимается только в таком смысле без дополнительных объяснений .

Схема работы системы распознавания следующая. Из сигнала выделяются участки, соответствующие фонемам ***к,п,с,т,ф,х,ц,ч,ш,щ,р*** которые классифицируются и маркируются символами “П”, “Ш”,”Р”. Затем сегментируются оставшиеся участки, отвечающие голосовым фонемам. Наконец, полученное разбиение и маркировка участков последовательно сопоставляются с транскрипцией каждого слова словаря. Вычисляется расстояние каждого отрезка до ближайшего эталона соответствующей фонемы. Сумма этих расстояний определяет расстояние сказанного до рассматриваемого слова словаря. Результат распознавания – слово, для которого расстояние минимально.

2. Выделение и классификация участков сигнала, соответствующих фонемам к,п,с,т,ф,х,ц,ч,ш,щ

Сигнал обрабатывается фильтром низких частот с частотой среза 500 Гц. На каждом отрезке разбиения вычисляется величина

$$E_{cp} = \sqrt{\frac{\sum_{i=1}^{256} |x_i - 127|^2}{256}} \qquad (1)$$

где x_i - значение i-го отсчёта отрезка. Далее вычисляется среднее значение величин (1), соответствующих всем отрезкам (порог). Тем отрезкам, для которых величина (1) больше ***половины*** среднего присваивается символ «В» , остальным символ «Н». Чтобы устранить случайные единичные включения для каждого i-го элемента полученной символьной последовательности S выполняется две обработки:

1-я обработка:

если S[i-1]=S[i+1] и S[i]≠S[i-1], то полагается S[i]=S[i-1]. (2)

2-я обработка:

если S[i]=S[i+3] и S[i+1]≠S[i], S[i+2]≠S[i] то полагается S[i+1]=S[i] и S[i+2]=S[i]. (3)

Считается, что совокупность идущих подряд отрезков, которым соответствует символ «Н», образуют участок, отвечающий одной из фонем к,п,с,т,ф,х,ц,ч,ш,щ. Остальные участки соответствуют голосовым фонемам.

Артикуляция фонем ***к,п,т*** связана с кратковременным перекрытием голосового тракта. Поэтому соответствующие участки речевого сигнала имеют малую амплитуду и их естественно квалифицировать как паузы. Для различения между собой шипящих и пауз используются специальные признаки: количество точек строгого максимума и количество точек постоянства. Точкой постоянства мы называем точку, в непосредственном соседстве с которой найдется другая точка, где сигнал принимает то же значение, что и в данной. Для шипящих звуков характерно малое число точек постоянства и большое число точек строгого максимума. Для пауз (при 8-битной записи, которую мы используем) – наоборот малое число точек строгого максимума и большое число точек постоянства. Для каждого из отрезков, отвечающих одной из фонем к,п,с,т,ф,х,ц,ч,ш,щ подсчитывается число точек постоянства ***с*** и число точек строго максимума ***m***. Если величина

$$c\text{-}m \qquad (4)$$

положительна, отрезок считается соответствующим паузе (программа маркирует участок паузы символом ”***П***”). Если величина (4) отрицательна, отрезок считается соответствующим шипящей (программа маркирует участок шипящей символом “***Ш***”). Каждой из фонем ***с,ш,щ*** соответствует ***Ш***-участок, для африкат ***ц,ч*** возможны сочетания ***Ш*** либо ***ПШ***. Для звуков ***ф,х*** возможны сочетания ***П*** и ***Ш*** в любом количестве и любом порядке. На рисунке 1 изображен сигнал, соответствующий слову “ОТСТАТЬ”

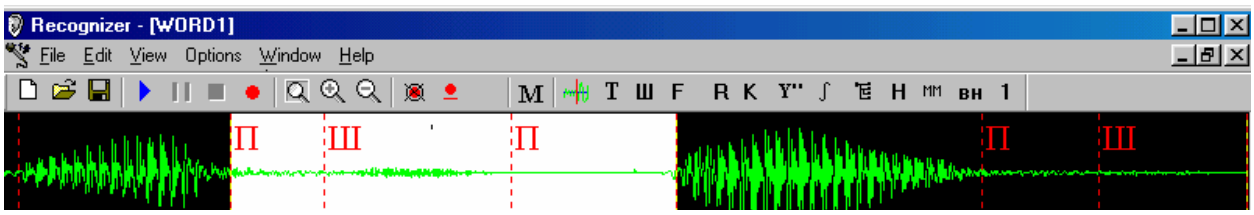

Рис. 1

На рисунке 2 представлен результат обработки участка «пауза – шипящая - пауза» этого сигнала с помощью вышеописанного алгоритма. В результате машина проставила метки, показанные на рисунке 1.

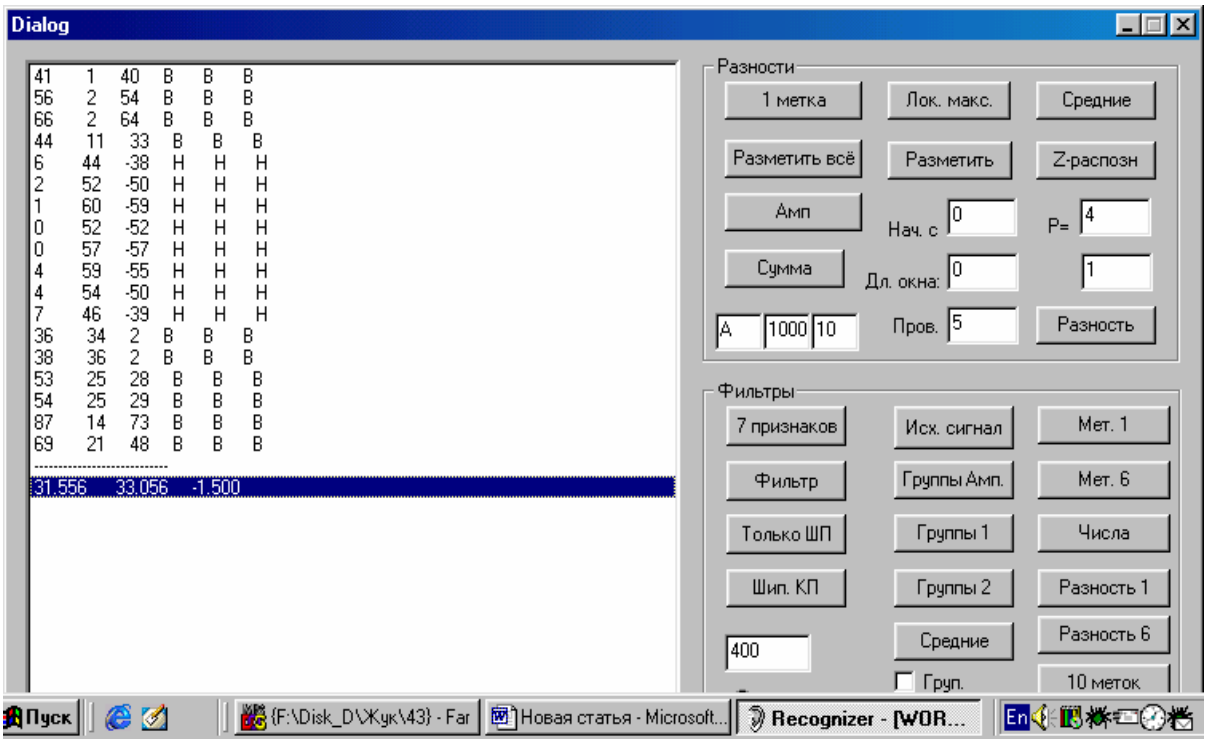

Рис.2

3. Выделение участков, отвечающих фонеме “Р”

На рисунке 3 изображен сигнал, соответствующего слову “ПАРА”. В нем выделен участок, соответствующий звуку “р” . Обработка происходит по следующему алгоритму (см. рисунок 4).

Сигнал, обработанный полосовым фильтром 300 – 1300 разбивается на отрезки, для которых вычисляются значения величины (1) (1-ый столбец на рисунке 4).

Условия проставления метки:

1 столбец : X[i] или X[i-1] или X[i+1] –й элемент- локальный минимум.

2 столбец (X2[i]=(X[i-1]-X[i])+(X[i+1]-X[i]))>20.

3 столбец (X3[i]=X2[i]-X[i])>-7 или X3[i-1]>-7 или X3[i+1]>-7.

4 столбец (X4[i]=X[i-1]-X[i])>3 или X4[i-1]>3 или X4[i+1]>3.

5 столбец (X5[i]=X[i+1]-X[i])>3 или X5[i-1]>3 или X5[i+1]>3.

Проверка квазипериодами: если в окрестностях предполагаемой метки происходит изменение длины квазипериода то метка считается правильной.

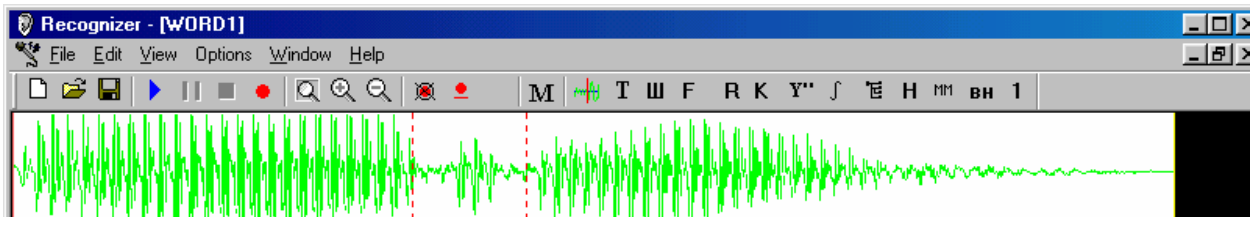

Рис. 3

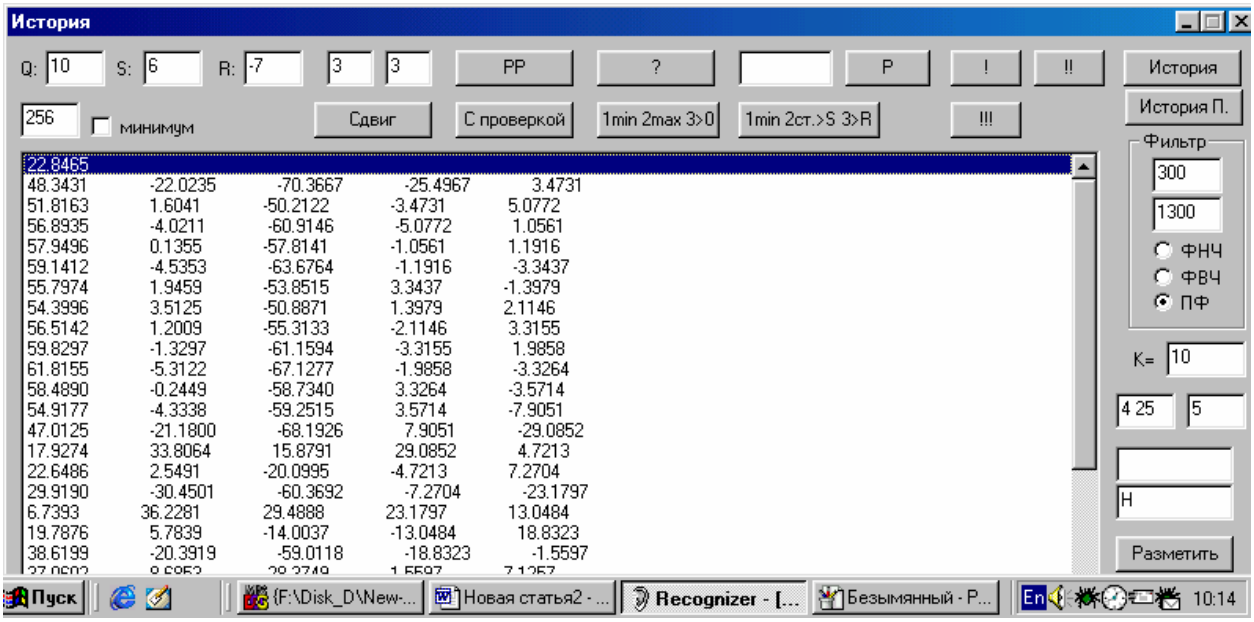

Рис. 4

4. Сегментация голосовых участков сигнала

Участок обрабатывается полосовым фильтром с полосой частот 300-1300 Гц, разбивается на отрезки по 256 отсчетов и на каждом отрезке вычисляется величина (1). Далее от начала участка последовательно берется m таких отрезков и вычисляется среднее значение соответствующих величин (1). Тем отрезкам, для которых величина (1) больше среднего присваивается символ «В» (выше порога), остальным символ «Н» (ниже порога). Чтобы устранить случайные единичные включения выполняются обработки (2) и (3).

Числу m последовательно придаются значения от 2 до 25. В результате получается набор последовательностей из «Н» и «В» длиной от 2 до 25. На рисунке 4 они ориентированы вертикально. Этот этап сегментации выполняется с использованием простейшей априорной транскрипции – “***а-м***”-транскрипции (см. п.6), отражающей чередование гласных и согласных звуков. При переходе от гласной (транскрипция “***а***”) к согласной (транскрипция “***м***”) метка ставится в той позиции N, в которой во всех последовательностях с длиной большей либо равной N все символы есть «Н». При переходе от согласной (транскрипция “***м***”) к гласной (транскрипция “***а***”) метка ставится в той позиции N, в которой во всех последовательностях с длиной большей либо равной N все символы есть «В».Таким образом, позиция метки в сигнале будет N*256 отсчетов. Далее вся описанная процедура повторяется, но началом отсчета теперь служит позиция предыдущей найденной метки.

На рисунке 5 изображен результат сегментации сигнала, соответствующего слову “ГАЗЕТА”. Рисунок 6 является образом окна, в котором наглядно представлен результат описанного поиска метки между участками “Г” и “А”.

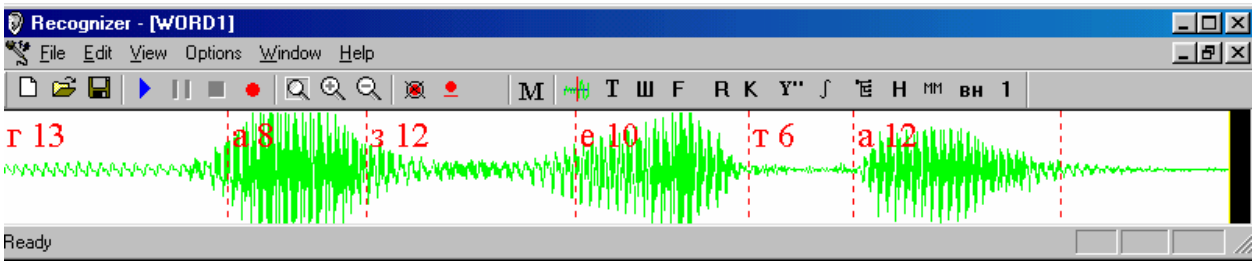

Рис. 5

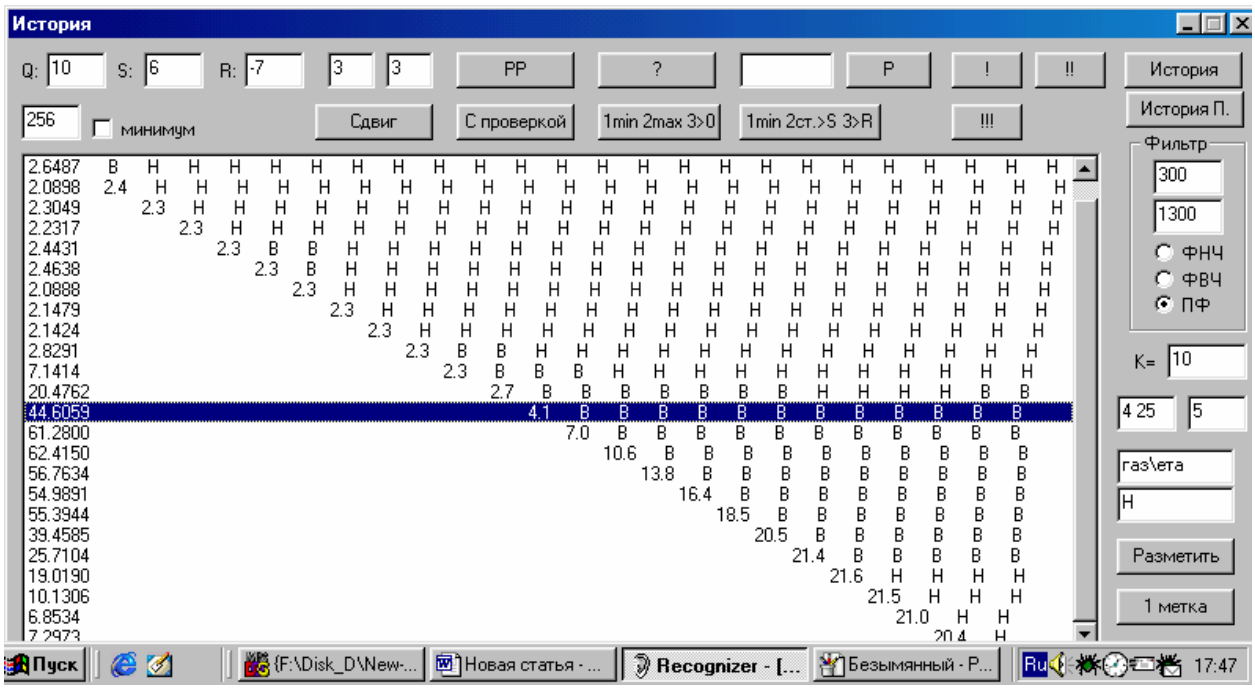

Рис. 6

5. Сегментация нескольких идущих подряд согласных

“Н-участок” между двумя гласными, выделенный на предыдущем этапе, может на самом деле содержать более одной согласной. Эксперимент показал, что в процессе естественной речи участки идущих подряд согласных имеют практически одинаковую длину. Проартикулировать две соседние согласные разной длины можно только в результате специальных усилий. Поэтому для того, чтобы выделить участки, отвечающие отдельным согласным мы, ориентируясь на транскрипцию, просто делим его на соответствующее число равных частей. На рисунке 5 показан результат сегментации слова “АЛМАЗ”.

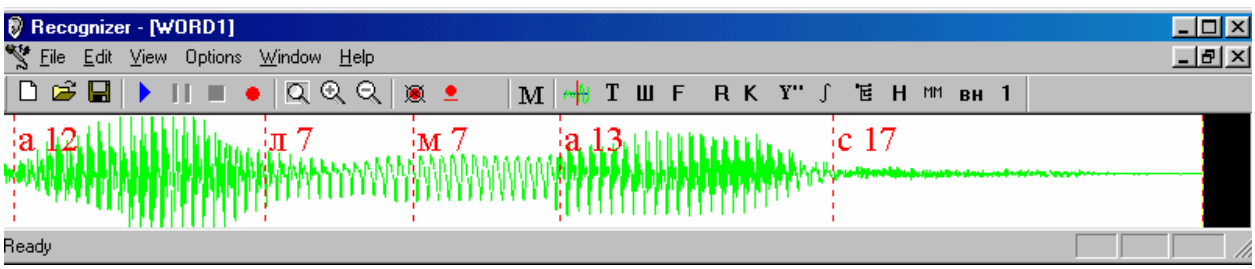

Рис. 5

6. Распознавание слов наперед заданного словаря

Для каждого слова словаря последовательно строится несколько видов транскрипции. Первоначально слова загружаются с обобщенной транскрипцией, содержащей лишь пять символов:

v - для группы идущих подряд голосовых кроме "р",
р для р,
ш - для шипящих;
п - для к,п,т;
х - для "ф", "х";

Слова, содержащие аффрикаты "ц","ч" снабжаются двойной транскрипцией. В первой указанные фонемы транскрибируются как ***ш***, во второй - как ***пш*** .

Машина с помощью описанных выше алгоритмов выделяет в сигнале “***ш, п, р*** – участки”. Остальные участки маркируются как ***v***. Для дальнейшей обработки пропускаются только слова, для которых транскрипция совпадает с последовательностью символов, маркирующих выделенные участки. При этом программой предусмотрено, что ***х*** может соответствовать любая последовательность ***ш,п*** - участков.

Далее слова, пропущенные для дальнейшей обработки снабжаются более подробной транскрипцией, где помимо указанных выше транскрипционных символов используются символы:

а - для гласных
м - для голосовых согласных.

Участки, ранее выделенные как голосовые, сегментируются с "***а-м***"-транскрипцией. При этом варианты сегментации, содержащие участки менее четырех и более двадцати пяти отрезков, считаются неверными и соответствующие слова отбрасываются.

Как показывает опыт, уже в результате описанной обработки отбрасывается не менее 90 процентов слов исходного словаря. Остальные слова снабжаются полной фонетической транскрипцией и для каждой из выделенных согласных вычисляется следующий функционал: число точек постоянства для участка сигнала делится на число точек постоянства для того же участка сигнала, обработанного фильтром низких частот с частотой среза 400 Гц. Если полученное отношение не меньше 4-х, то компьютер считает, что этот участок соответствует "шумной" фонеме "ж" или "з", в противном случае - одной из остальных голосовых согласных. Соответственно отбрасывается еще некоторое количество слов.

Если оставшийся после отбрасывания “лишних” слов список содержит лишь одно слово, то оно считается результатом распознавания. В противном случае для каждого из оставшихся слов вычисляется некоторая функция расстояния. При этом используются эталоны фонем, созданные в процессе обучения конкретным диктором. Каждому транскрипционному символу при этом оказываются соответствующими несколько эталонов, построенных с использованием признаков, о которых будет сказано ниже. Машина, уже выполнив к этому моменту сегментацию, соотносит каждый отрезок с соответствующим транскрипционным символом. Далее вычисляется расстояние до всех эталонов соответствующей фонемы и выбирается наименьшее из них. Все эти расстояния

суммируются и полученная сумма считается расстоянием до соответствующего слова. Результатом распознавания объявляется слово, расстояние до которого минимально.

7. О системе признаков, используемых при создании эталонов и распознавании фонем.

Мы используем три системы признаков:

1. Дельта-представление логарифма кумулятивного отношения (см. [3]).
2. Результат обработки кумулятивного отношения методами аналогового декодирования.
3. Вектор признаков, использующий коэффициенты линейного предсказания (см. [4]).

Соответственно создается три системы эталонов и вычисляется три расстояния сказанного до каждого слова заключительного списка, допущенного к распознаванию. Результат распознавания определяется голосованием: слово должно быть распознанным как минимум двумя системами. Если результаты всех трех систем различны, то программа констатирует отказ от распознавания.

Литература

1. Дорохин О.А., Старушко Д.Г., Федоров Е.Е., Шелепов В.Ю. Сегментация речевого сигнала // Искусственный интеллект. – 2000. - №3. – С.450-458.

2. Шелепов В.Ю., Ниценко А.В. Амплитудная сегментация речевого сигнала, использующая фильтрацию и известный фонетический состав.

3. .Старушко Д.Г., Шелепов В.Ю. Новая система признаков для распознавания речевых единиц. Искусственный интеллект. – 2002.- №4.- С.286-288

4. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. – М.: Радио и связь, 1981. – 495 с.