а) Показать, что погрешность предсказания

$$E^{(p)} = \sum_{n=0}^{N-1+p} e^2(n) = \sum_{n=0}^{N-1+p} \left[ -\sum_{i=0}^{p} \alpha_i x(n-i) \right]^2$$

может быть записана в виде $E^{(p)} = \boldsymbol{\alpha} \mathbf{R}_{\alpha} \boldsymbol{\alpha}^t$, где $\mathbf{R}_{\alpha}$ — матрица $(p+1) \times (p+1)$. Определить $\mathbf{R}_{\alpha}$.

б) Предположим, что сигнал $\hat{x}(n)$ пропущен через обратный фильтр с коэффициентами $\alpha$, что дает погрешность предсказания $\hat{e}(n)$, определяемую выражением

$$\tilde{e}(n) = -\sum_{i=0}^{p} \alpha_i \hat{x}(n-i).$$

Показать, что средняя квадратическая ошибка $\tilde{E}^{(p)}$, определяемая как $\tilde{E}^{(p)} = \sum_{n=0}^{N-1+p} [\tilde{e}(n)]^2$, может быть записана в виде $\tilde{E}^{(p)} = \mathbf{d} \underset{\alpha}{\mathbf{R}} \boldsymbol{\alpha}^t$, где $\underset{\alpha}{\mathbf{R}}$ — матрица $(p+1) \times (p+1)$. Определить $\underset{\alpha}{\mathbf{R}}$.

в) Если определить отношение $D = \tilde{E}^{(p)}/E^{(p)}$, то что можно сказать о диапазоне значения $D$?

**8.12.** Предложена следующая мера различимости между двумя сегментами речевого сигнала с параметрами предсказания $\boldsymbol{\alpha}$ и $\hat{\boldsymbol{\alpha}}$ и корреляционными матрицами $\mathbf{R}_{\alpha}$ и $\underset{\alpha}{\mathbf{R}}$ (см. задачу 8.11):

$$D(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) = \frac{\boldsymbol{\alpha} \underset{\alpha}{\mathbf{R}} \boldsymbol{\alpha}^t}{\hat{\boldsymbol{\alpha}} \underset{\alpha}{\mathbf{R}} \hat{\boldsymbol{\alpha}}^t}.$$

а) Показать, что мера различимости $D(\alpha, \hat{\alpha})$ может быть записана в следующей удобной для вычислений форме:

$$D(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) = \left[ \frac{\left( b(0) \hat{R}(0) + 2 \sum_{i=1}^{p} b(i) \hat{R}(i) \right)}{\hat{\boldsymbol{\alpha}} \underset{\alpha}{\mathbf{R}} \hat{\boldsymbol{\alpha}}^t} \right],$$

где $b(i)$ — автокорреляционная функция вектора $\boldsymbol{\alpha}$ — равна:

$$b(i) = \sum_{j=0}^{p-i} \alpha_j \alpha_{j+i}, \quad 0 \leqslant i \leqslant p.$$

б) Предположим, что величины (вектора, матрицы, скаляры) $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}$, $\mathbf{R}$, $\hat{\mathbf{R}}$, $(\alpha \underset{\alpha}{\mathbf{R}} \hat{\boldsymbol{\alpha}}^t)$, $\underset{\alpha}{\mathbf{R}}$ и $\mathbf{b}$ вычислены заранее, т. е. известны к моменту расчета меры расстояния. Сравнить объем вычислений, необходимый для определения $D(\alpha, \hat{\alpha})$, используя оба выражения для $D$, рассмотренные в данной задаче.



# 9

# Цифровая обработка речи в системах речевого общения человека с машиной[1]

## 9.0. Введение

В предыдущих главах внимание было сконцентрировано на основных теоретических вопросах, необходимых для понимания современных методов цифровой обработки речевых сигналов. Еще не рассматривалась обширная область применения разработанных методов, т. е. способов использования моделей и связанных с ними параметров в системах передачи или автоматического выделения информации из сигнала речи. В данной главе приведены характерные примеры цифровой обработки речи применительно к системам общения между человеком и машиной (ЭВМ) посредством голоса. Существует ряд причин, по которым имеет смысл ограничиться рассмотрением примеров связи между человеком и машиной. Прежде всего, эта область наиболее плодотворна с точки зрения возможностей использования методов цифровой обработки речи и позволяет, таким образом, проиллюстрировать почти все рассмотренные выше методы обработки. Кроме того, эта область является чрезвычайно важной, дающей все новые и новые приложения, область, которая только еще развивается и демонстрирует огромные возможности для широкого применения.

Системы речевого обмена между человеком и машиной можно подразделить на три класса: с речевым ответом, распознавания диктора и распознавания речи.

Системы с речевым ответом предназначаются для выдачи информации пользователю в форме речевого сообщения. Таким образом, системы с речевым ответом — это системы односторонней связи, т. е. от машины к человеку. С другой стороны, системы второго и третьего классов — это системы связи от человека к машине. В системах распознавания диктора задача состоит в верификации диктора (т. е. в решении задачи о принадлежности данного диктора к некоторой группе лиц) или идентификации диктора из некоторого известного множества. Таким образом, класс задач распознавания диктора распадается на два подкласса: верификации и идентификации говорящего. Различия и сходство между этими задачами будут рассмотрены в последующем.

Последний класс задач распознавания речи также можно разделить на подклассы в зависимости от таких факторов, как размер словаря, количество дикторов, условия произнесения слов и т. д. Основная задача распознающей системы сводится либо к точному распознаванию произнесенной на входе фразы (т. е. система фонетической или орфографической печати произнесенного текста), либо к «пониманию» произнесенной фразы (т. е. к правильной реакции на сказанное диктором). Именно задача понимания, а не распознавания наиболее важна для систем с достаточно большим словарем непрерывных речевых сигналов, в то время как задача точного распознавания более важна для систем с ограниченным словарем, малым количеством дикторов, систем распознавания изолированных слов. Различные аспекты построения систем распознавания речи также рассматриваются в данной главе.

В заключительной части главы рассмотрены некоторые системы, типичные для каждой области речевого общения человека с машиной. Более подробно рассматриваются особенности обработки речевого сигнала с целью закрепления результатов предшествующих глав. Однако как для полноты обсуждения, так и для более глубокого понимания здесь излагаются и общие аспекты обработки информации в системе, поскольку часто они оказываются достаточно важными для успешной работы системы в целом.

[1] Имеются в виду цифровые ЭВМ. (Прим. ред.)

## 9.1. Системы с речевым ответом

На рис. 9.1 представлена общая структурная схема системы с речевым ответом на базе ЭВМ. Элементами этой системы являются блоки: памяти для хранения словаря системы с речевым ответом; хранения правил синтеза сообщений по элементам словаря; программ формирования речевого ответа.

На вход системы с речевым ответом поступает сообщение о содержании вопроса, порождаемого либо другой системой обработки информации, либо непосредственно от человека, обратившегося с интересующим его вопросом к информационной системе.

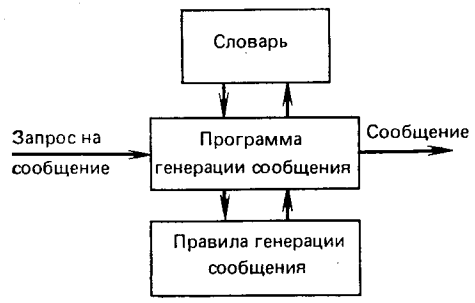


Рис. 9.1. Структурная схема системы с речевым ответом

Откликом системы на поставленный вопрос служит выходное сообщение в виде речевой фразы. Простым примером такой системы является автоматическая справочная телефонная служба, которая обнаруживает неправильно набранный номер, определяет причину ошибки (например, телефон отключен или ему присвоен новый номер и т. д.) и посылает на выход системы с речевым ответом сообщение, содержащее необходимые абоненту указания. В таких системах словарь обычно состоит из ограниченного набора изолированных слов (например, цифр с различными окончаниями).

В качестве другого примера рассмотрим информационную систему о состоянии курса акций. Здесь абонент должен с помощью кнопочного набора ввести код интересующего его курса. Система декодирует набор, определяет текущий курс акций и затем выдает соответствующую информацию в систему с речевым ответом для составления требуемой фразы. В данном случае словарь должен содержать достаточно широкий набор различных слов и фраз.

Существуют два основных подхода к построению систем с речевым ответом. Один из них заключается в попытке построения системы, речевые возможности которой сравнимы с возможностями человека. Такие системы (называемые часто системами синтеза речи по правилам) основаны на модели речеобразования, рассмотренной в гл. 3. В этом случае для синтеза достаточно хранить словарь произношений элементов. Сигналы, необходимые для управления речевым синтезатором, в соответствии с моделью речеобразования формируются на основе правил синтеза. Такие системы представляют интерес в том случае, если требуется словарь весьма большого объема. Реализация подобных систем — это проблема, требующая чрезвычайно трудоемких исследований, и на этапе синтеза сигнала имеются обширные возможности применения рассмотренных выше методов цифровой обработки сигналов. Однако



основная трудность при построении подобных систем состоит в разработке правил управления синтезатором. В данной книге примеры таких систем не рассматриваются, поскольку это увело бы нас в область лингвистики. Интересующихся читателей отсылаем к работам [2—6].

В системах с речевым ответом второго типа используется ограниченный словарь и сигнал на выходе таких систем формируется посредством сочленения отдельных элементов реального речевого сигнала, взятых из словаря. На рис. 9.2 представлена структурная схема системы, в которой словарь, состоящий из отдельных

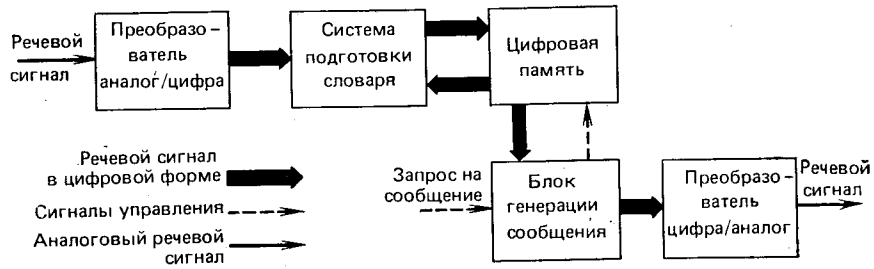


Рис. 9.2. Структурная схема системы с речевым ответом [8]

слов, представленных в цифровой форме, хранится в памяти. Сообщения конструируются в этом случае путем отыскания требуемых слов и фраз в памяти и воспроизведения их в требуемой последовательности. При разработке систем подобного типа следует учитывать три основных соображения. Во-первых, способ представления и хранения словаря должен быть выбран таким образом, чтобы в разработанной системе имелась возможность свободного доступа к любому элементу словаря. Во-вторых, должен быть выбран способ редактирования речевого материала словаря совместно со способом записи его элементов в память. В-третьих, необходимо обеспечить заданную последовательность выбора и воспроизведения элементов словаря (т. е. способ формирования сообщения).

Поскольку назначение систем с речевым ответом состоит в формировании речевых сообщений, предназначенных для человека, требование к разборчивости становится определяющим. Не менее важное значение, однако, имеют и такие параметры речи, как качество восприятия и натуральность. Таким образом, в разрабатываемой системе необходимо с предельной полнотой реализовать все три основных условия с тем, чтобы добиться максимально возможной разборчивости и натуральности речевого сигнала.

### 9.1.1. Основные аспекты построения систем с речевым ответом

Развитие методов цифрового представления и цифровой обработки сигналов, а также методов построения цифровых устройств

позволяет создавать системы речевого ответа, выполненные полностью на базе цифровой техники. В показанной на рис. 9.2 цифровой системе необходимо, прежде всего, осуществить аналого-цифровое преобразование, т. е. представить речевой сигнал в цифровой форме. Аналогично для преобразования цифрового представления в аналоговую форму требуется цифроаналоговый преобразователь. Поскольку словарь представлен в цифровой форме, его можно хранить в цифровой памяти. Для доступа к элементам словаря в нужной последовательности и составления из них требуемой фразы необходима система формирования сообщений. Полученное цифровое представление синтезированной фразы, в свою очередь, поступает на цифроаналоговый преобразователь.

Центральным фактором, определяющим сложность систем с речевым ответом, является выбор способа цифрового представления речи при составлении словаря. Как будет ясно из дальнейшего, здесь имеются широкие возможности использования различных способов цифрового представления, начиная со способов кодирования речевых колебаний (см. гл. 5) и кончая системами «анализ—синтез» (см. гл. 6—8). Выбор способа цифрового представления оказывает большое влияние на объем и тип цифровой памяти, а также на способ синтеза речевого сообщения.

При рассмотрении способа цифрового представления речевого сигнала применительно к системам с речевым ответом полезно остановиться на трех основных моментах:

скорость передачи информации (в битах в секунду), необходимая для получения приемлемого качества;

сложность способа кодирования и декодирования;

гибкость представления, т. е. возможность модификации элементов словаря.

На рис. 9.3 показаны результаты сравнительного анализа методов цифрового представления, рассмотренных в гл. 5—8, по трем


Рис. 9.3. Обзор методов кодирования речи

перечисленным выше показателям. Представление на основе кодирования речевого колебания требует наибольших скоростей передачи и, следовательно, максимального объема памяти для хранения элементов словаря. Эти способы являются простейшими с точки зрения алгоритмов кодирования—декодирования. С другой стороны, способы анализа—синтеза, которые буквально «разбирают речевой сигнал на части», обладают широкими возможностями полезной модификации элементов словаря. Два первых фактора, т. е. скорость передачи и сложность реализации, оказывают существенное влияние на технико-экономические показатели при разработке полностью цифровых систем речевого ответа. Рассмотрим типовой словарь, объем которого составляет 100 слов со средней протяженностью 1 с. В табл. 9.1 показаны оценки (весьма осторожные) объема памяти, необходимого для хранения произноси-



Таблица 9.1

**Объем памяти, необходимой для хранения цифрового представления речи**

| Метод кодирования | Скорость, кбит/с | Объем памяти, бит | Примерная стоимость, долл. |
|---|---|---|---|
| ИКМ | 40 | 4 000 000 | 4000 |
| АРИКМ | 24 | 2 400 000 | 2400 |
| Линейное предсказание | 2,4 | 240 000 | 240 |
| Форманты | 0,6 | 60 000 | 60 |

мой в течение 100 с речи. Даже при использовании для кодирования логарифмической ИКМ объем памяти остается вполне приемлемым. Основной вопрос заключается в соотношении между стоимостями цифровой памяти и аппаратуры кодирования—декодирования для наиболее сложных систем кодирования. В последнем столбце табл. 9.1 приведена стоимость памяти системы речевого ответа из расчета 0,1 цента за бит[1]. С учетом того, что один блок памяти может быть использован для ряда каналов, стоимость устройства кодирования—декодирования должна быть достаточно малой с тем, чтобы не являться определяющей частью общей стоимости системы. Совершенно очевидно, например, что стоимость формантного синтезатора, необходимого для высококачественного воспроизведения речевого сигнала, окажется значительно большей, чем стоимость соответствующего блока памяти для словаря малого объема.

Другой важной задачей, решаемой при построении систем с речевым ответом, являются создание и редактирование словаря. При решении этой задачи, т. е. подготовке элементов словаря и обеспечении высококачественного сигнала на выходе, цифровые методы оказываются чрезвычайно эффективными и гибкими. Обычно слова и фразы, включаемые в словарь, произносятся специально обученным диктором и записываются с высоким качеством. Затем слова или фразы подвергаются аналого-цифровому преобразова-

[1] Это довольно грубая верхняя граница действительной стоимости одного бита памяти.

нию и кодированию. Цифровое представление (которое может быть как описанием формы сигнала, так и основанным на представлении типа «анализ—синтез») оперативно хранится в цифровой форме в ЭВМ. Для исключения пауз между фразами используется специальный метод поиска начала и конца фразы. Как показано в гл. 4, при высококачественной записи начало и конец каждой фразы можно определить с высокой точностью. При этом можно точно сказать, удовлетворяет ли протяженность данной фразы заданной. Фраза, кроме того, может быть воспроизведена для проверки окончаний слов или фразы на слух. Записи можно легко повторять, пока не будут достигнуты требуемые длительность и окончание вводимой фразы.

Заключительным шагом в создании словаря являются сравнение энергетических уровней всех слов в словаре и соответствующее изменение уровней для получения некоторого единого уровня или такого распределения уровней, которое предопределяется предполагаемым использованием словаря. Это может быть сделано или на основе вычисления максимального значения сигнала, или на основе использования других мер, таких, как кратковременная энергия.

Если слово или фраза записаны с требуемым качеством, то они хранятся в определенном месте памяти словаря. Это достигается простой установкой файлов в речевой системе и указанием адресов, которые используются системой синтеза фраз для определения начала и окончания каждого элемента словаря.

Помимо рассмотренных методов создания словаря система с речевым ответом включает в себя методы синтеза фраз по элементам словаря. В этом случае методы цифрового представления также обладают значительными преимуществами. Если используется метод кодирования формы речевого колебания, то все, что здесь необходимо, — это сочленить речевые сигналы элементов словаря. Если элементом словаря является отдельное слово, то такой метод может привести к некоторой потере натуральности звучания, но подобный подход обладает важным преимуществом, состоящим в том, что система синтеза фраз оказывается очень простой. В самом деле, такая система легко может быть выполнена на основе микропроцессора. Пример подобной системы рассматривается в 9.1.2.

С другой стороны, представление, основанное на преобразовании типа «анализ—синтез», обладает большой гибкостью по отношению к изменяющимся свойствам элементов словаря, например временным соотношениям, окончаниям и т. д. Это свойство является даже более важным, чем малая скорость передачи (объем описания), которую можно достигнуть при использовании описания на основе преобразования «анализ-синтез». Поскольку элементы словаря представлены в виде набора основных параметров речевого сигнала, можно, например, изменять период основного тона и длительность слов таким образом, чтобы привести их в соответствие с контекстом. Более интересной представляется воз-



можность такого изменения параметров на границах слов, чтобы добиться как можно большего сходства между синтезированными и реальными речевыми сигналами. Достигнуть такого эффекта даже в простейших случаях можно лишь на основе использования правил для определения требуемого периода основного тона и протяженности во времени, а также алгоритмов изменения параметров в соответствии с изменяющейся протяженностью слов и поглощением их границ в слитной речи. Вследствие малого объема параметрического описания словаря для получения удовлетворительного качества синтезированного сигнала системы с речевым ответом на основе преобразования «анализ—синтез» следует реализовывать аппаратурно с использованием микропроцессоров. Пример системы такого рода рассмотрен в 9.1.3.

### 9.1.2. Многоканальная цифровая система с речевым ответом

На рис. 9.4 показана структурная схема многоканальной цифровой системы, созданная в лабораториях Белла с применением малого спецвычислителя [7, 8]. В этой системе элементы словаря представлены с помощью АРИКМ со скоростью 24 кбит/с. Кодер и декодеры АРИКМ реализованы в виде отдельных устройств.

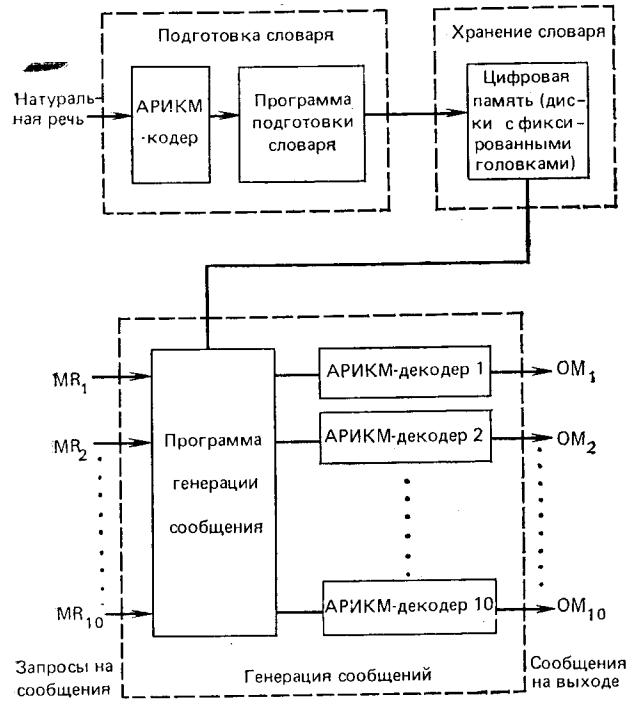


Рис. 9.4. Структурная схема многоканальной системы с речевым ответом [7]

Начало и конец каждого слова определяются автоматически с использованием алгоритма, основанного на вычислении кратковременной «энергии» АРИКМ кодовых слов [7]. Словарь хранится на диске с фиксированным положением головок для быстрого поиска элементов словаря по командам программы синтеза фраз. Эта часть системы выполняет, главным образом, логические операции и операции пересылки данных. Важной особенностью, свойственной многим вычислительным машинам и системам памяти, является возможность одновременного обслуживания ряда информационных каналов единой словарной памятью. Синтезированные в ЭВМ фразы запрашиваются другими компьютерами либо по телефонным линиям, либо непосредственно через цифровые устройства ввода—вывода. Программа синтеза сообщений определяет необходимые элементы словаря и пересылает их в память. Цифровое представление требуемых сообщений хранится в буферной памяти машины со свободным прямым доступом. Буферные устройства связаны с соответствующими АРИКМ декодерами с помощью каналов прямого доступа в память. При такой организации система обеспечивает одновременное обслуживание ряда каналов. В лабораториях Белла создана десятиканальная система с речевым ответом. Она применяется в ряде приложений, как это рассмотрено в 9.1.4.

### 9.1.3. Система синтеза речи на основе последовательного объединения слов, закодированных формантами

В качестве примера использования преобразования типа «анализ—синтез» рассмотрим структурную схему, представленную на рис. 9.5. В этом случае элементы словаря сформированы так, как это описано в § 7.4, где речевой сигнал представлен набором параметров, например период основного тона, признак вокализованный — невокализованный, интенсивность и формантные частоты. Таким образом, для хранения слов и фраз словаря достаточно объема памяти 600 бит, приходящейся на 1 с длительности сигнала.

Помимо элементов словаря система синтеза сообщений должна включать методы обработки, обеспечивающие получение требуемой длительности слов и периода основного тона в синтезируемой фразе. На основе этих данных формантные частоты соседних слов сглаживаются так, чтобы они непрерывно переходили одна в другую аналогично слитной речи. На рис. 9.6 показано, как использование формантного описания позволяет преобразовывать элементы словаря для достижения натуральности звучания. Длительность можно изменять посредством интерполяции. Кроме того, на стыке второго и третьего слов формантные частоты подстраиваются таким образом, чтобы их траектории не разрывались при переходе через границу, как это и должно быть в слитной речи. Заметим, наконец, что контуры основного тона каждого из элементов словаря могут быть изменены или опущены для получения единого контура основного тона, отвечающего фразе в целом.



На рис. 9.7 представлен пример [10], иллюстрирующий различие между сочленением осциллограмм речевого сигнала и сочленением слов, представленных формантами. На рис. 9.7*а* показана

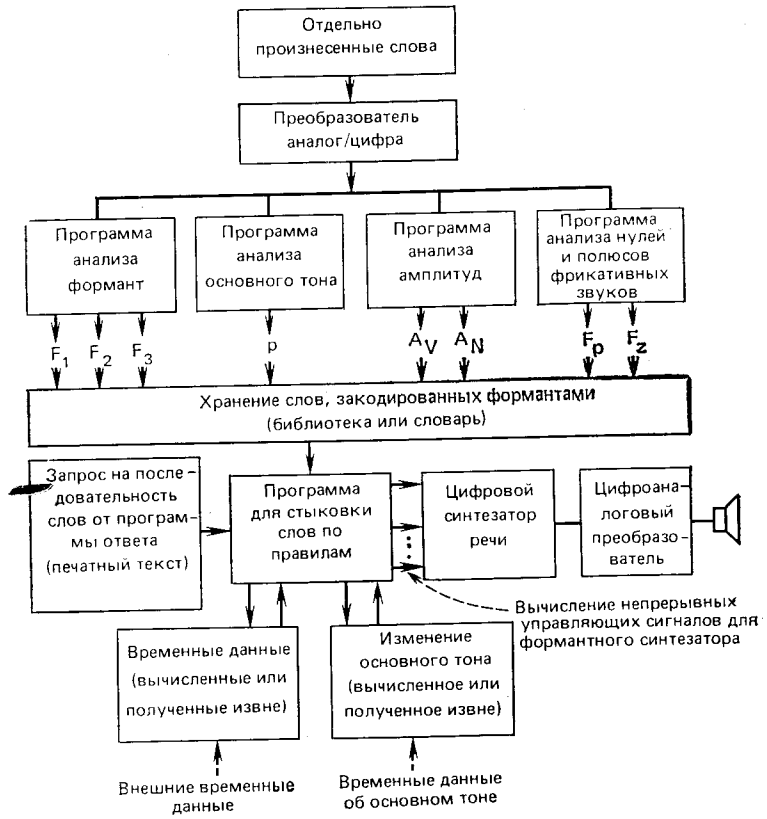


Рис. 9.5. Структурная схема системы с речевым ответом, основанная на формантном представлении [9]

спектрограмма исходной фразы «I am an aspiring orator». На рис. 9.7*в* показана спектрограмма той же фразы, полученная путем сочленения отдельно произнесенных слов, закодированных формантами без изменения периода основного тона или их длительности. В данном случае видны разрывы в формантных траекториях. На рис. 9.7*б* показана спектрограмма фразы, полученной из последовательности слов, закодированных формантами путем соответствующего согласования частот на границах слов. Контур основного тона и длительность слов, показанных на рис. 9.7*б*, рассчитаны по данным, полученным в результате исследования системы синтеза речи по правилам, разработанным Кокером и

Умедой [3, 5, 6]. Длительности слов на рис. 9.7*а* и *б* вполне соизмеримы, а соответствующие формантные траектории весьма похожи друг на друга.

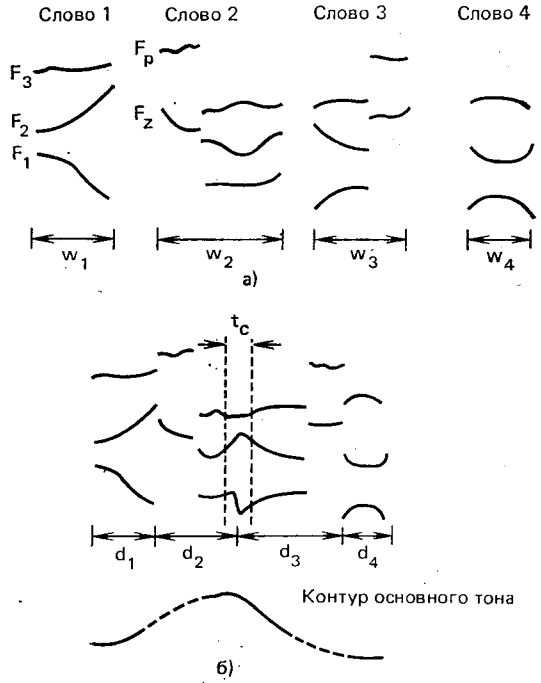


Рис. 9.6. Типичный пример, иллюстрирующий процесс генерации сигналов управления по словам из словаря. Представлен случай сообщения из четырех слов; все параметры — функции времени [9]

Система, изображенная на рис. 9.5, использовалась для синтеза телефонных сообщений вида «Номер 135-3201» [9]. Правила управления основным тоном и длительностью произнесения семизначной последовательности цифр подбирались эмпирически по измерениям параметров реального речевого сигнала. При этом оказалось, что синтезированная речь, формируемая системой, изображенной на рис. 9.5, является более предпочтительной по сравнению с речью, полученной путем простого сочленения последовательности слов. Это связано с наличием «машинного» акцента у синтезированной речи. Хотя эти результаты и являются обнадеживающими, однако для построения методики синтеза, приводящей к натурально звучащей высококачественной синтезированной речи, сформированной по словарю фраз или слов, представленных в цифровой форме, требуются обширные исследования [11].

В заключение отметим, что имеется целый ряд способов цифрового представления элементов словаря, позволяющих столь же гибко манипулировать с параметрами речевых фраз.



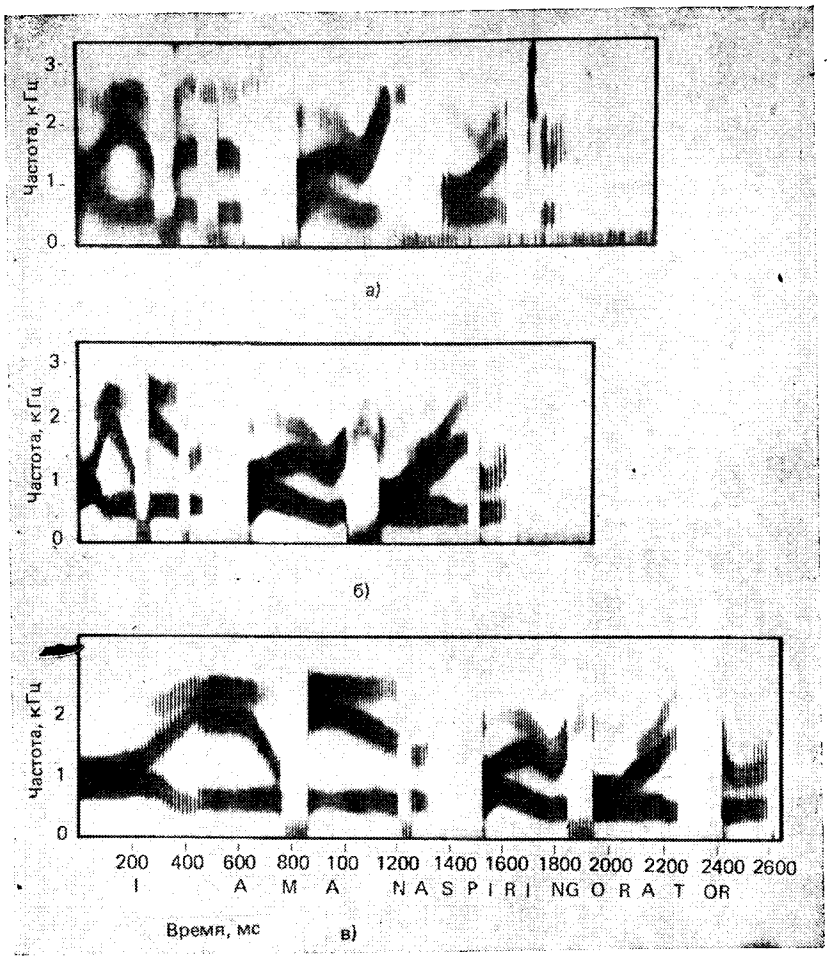


Рис. 9.7. Сравнение спектрограмм:
*а*) исходного сигнала; *б*) слов, модифицированных по просодике, периоду основного тона и временным соотношениям; *в*) последовательности изолированных слов [10]

### 9.1.4. Применение систем с речевым ответом

Гибкость систем с речевым ответом облегчает их использование в ряде экспериментальных работ, выполняемых в лабораториях Белла. В настоящее время созданы и исследованы: система речевых команд для выполнения межблочных соединений аппаратуры связи; вспомогательная справочная система; система информации о текущем курсе акций; система информации и контроля банков данных; справочная авиаслужба; система верификации дикторов. Ниже описываются две из перечисленных систем.

**Применение систем речевого ответа при производстве работ по монтажу оборудования электросвязи.** Обычно монтажник работает по отпечатанному тексту, содержащему указания по каждому межблочному соединению. Однако в ряде случаев при монтаже оборудования монтажнику бывает неудобно отрывать глаза от работы для чтения инструкции. В подобных случаях удобнее записать инструкцию на кассетный магнитофон и дать возможность монтажнику работать с указаниями в форме магнитофонной записи. Обычно для запуска магнитофона используют ножной выключатель, а останавливается он по тональному сигналу, записанному в конце каждого указания.

Таблица соединений может быть продиктована человеком. Однако для этого один человек должен прочитать указания (затратив на это, может быть, несколько часов), а другой — проверить запись во избежание ошибок. Обнаруженные неточности затем исправляются. Даже после успешной записи инструкции через некоторое время может потребоваться ее уточнение, что приведет к необходимости повторения всего процесса. Потребность в уточнении указаний может возникать несколько раз на протяжении нескольких дней или недель. Повышение утомляемости приводит к возрастанию числа ошибок, допускаемых людьми.

Таким образом, система с речевым ответом обладает рядом преимуществ:

1. Указания по монтажу обычно состоят из набора простых команд, содержащих лишь необходимую информацию, такую, как цвет и длина провода, точки его подключения. В этих случаях не требуется сглаженная или слитная речь.

2. Инструкции могут быть сформированы на основании относительно малого словаря — около 50 слов достаточно для части аппаратуры и около 100 слов — для комплекса оборудования связи, монтируемого фирмой «Вестерн Электрик».

3. Инструкции обычно формируются с помощью ЭВМ, поэтому их удобно использовать в системах с речевым ответом.

4. Инструкции часто изменяются. Использование систем с речевым ответом позволяет упростить утомительную работу по пересмотру содержания инструкций.

На рис. 9.8 представлена структурная схема системы с речевым ответом, предназначенная для формирования указаний по монтажу оборудования электросвязи. В качестве исходной информации для создания инструкции используется колода перфокарт, полученная с ЭВМ фирмы «Вестерн Электрик». Отперфорированные символы описывают слова, предназначенные для конкретной инструкции. Например, фраза

КРАСНЫЙ-ПАУЗА-20-7-ПАУЗА-4-Р-ПАУЗА-7-Z-КОНЕЦ

сообщает монтажнику, что красный провод длиной 27 дюймов следует пропустить от точки 4Р к точке 7Z. Используя соответствующий словарь, система с речевым ответом формирует нужное сообщение и посылает его к АРИКМ-декодеру. Полученный фрагмент речевого сообщения (указание монтажнику) записывается на кассету магнитофона.



В рассмотренном примере не используются возможности системы, связанные с передачей сообщений по каналу связи, однако можно в полной мере использовать достоинства многоканального

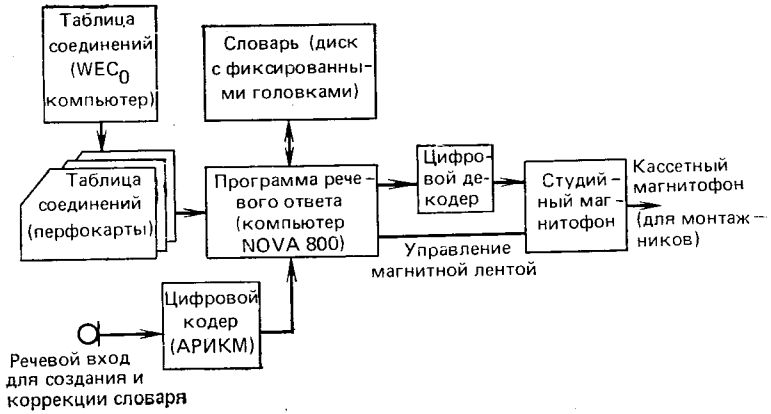


Рис. 9.8. Система речевого ответа для автоматического синтеза инструкции по монтажу [8]

режима работы как для одновременного формирования нескольких инструкций, так и для параллельного формирования одной слишком длинной инструкции с целью сокращения времени, затрачиваемого на ее запись [7, 8].

**Системы с изменяющимся содержанием информационного банка.** При производстве монтажных работ с применением систем речевого ответа между возможным пользователем и системой формирования сообщений нет взаимодействия. Это связано с тем, что кнопочный ввод в систему заменен предварительно отперфорированным набором карт, которые определяют сообщение, требуемое на выходе системы. При использовании систем с речевым ответом в качестве вспомогательной справочной службы, выдающей справки о кредитах, состоянии текущего счета, наличии товаров, необходим доступ к содержанию информационного банка. Система должна находить нужную информацию и формировать соответствующее сообщение, поступающее к абоненту. На рис. 9.9 представлена структурная схема системы с изменяющимся содержанием информационного банка. Здесь предпо-

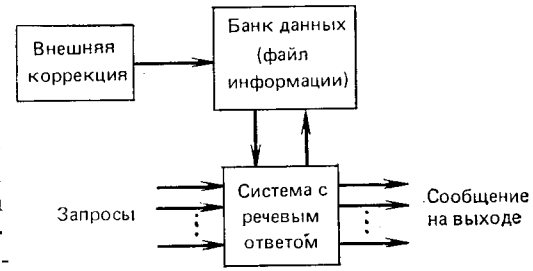


Рис. 9.9. Структурная схема системы речевого ответа с изменяющимся словарем [8]

лагается, что содержание банка данных может быть изменено как с помощью дополнительного внешнего источника, так и самой системой.

Предположим, что содержанием банка информации являются опись количества товаров, производимых компанией, и объем реализованной продукции с распределением по наименованиям. После каждой проведенной сделки информационная система с речевым ответом должна вносить изменения в содержание банка данных. По мере производства товаров банк данных также должен уточняться. В этом примере информационная система с речевым ответом не только позволяет вести учет товаров, но и предотвращает возможность продажи несколькими агентами одного и того же изделия в случае, когда их количество ограничено. Учет товаров также дает возможность компании всегда иметь текущую статистику спроса и, таким образом, корректировать ассортимент производимой продукции в соответствии со спросом.

Интересно также применение системы с коррекцией информационного банка в качестве системы информации о текущем курсе акций. Содержание банка данных составляют сведения о рыночной стоимости любой акции. Внешняя коррекция данных производится непосредственно с телеграфной или телетайпной ленты, содержащей последние биржевые новости.

Система информации о текущем курсе акций используется примерно таким образом. Абонент вызывает систему, которая отвечает: «Это система информации о текущем курсе акций. Стоимость приводится также и по отношению к ближайшему прошлому рабочему дню. Пожалуйста, введите рыночное обозначение интересующей вас акции». Абонент вводит: А-Т-Т-* — и система отвечает: «Американские Телефонные и Телеграфные, 62 и 3/8, вверх на 1/4».

Несомненно, что в будущем системы с речевым ответом на основе ЭВМ найдут широкое применение. Очевидно также, что ключевую роль при построении таких систем будут играть методы цифровой обработки речевых сигналов.

## 9.2. Системы распознавания дикторов

При распознавании дикторов цифровая обработка речи является тем первым шагом, с которого начинается решение задачи распознавания образов. Как видно из рис. 9.10, речевой сигнал (представление образа вектором) представлен с использованием таких методов цифровой обработки, которые сохраняют индивидуальные особенности диктора. Полученный образ сравнивается с предварительно подготовленными эталонными образами, а затем применяется соответствующая логика принятия решений для определения голоса заданного диктора среди возможного множества. Системы распознавания дикторов подразделяются на два вида: идентификация и верификация. При верификации диктора требуется установить его идентичность данному эталону. Устройство верификации принимает одно из двух возможных решений: диктор является тем, за кого он себя выдает, или не является. Для вынесения такого решения используется совокупность параметров, содержащих необходимую информацию об индивидуальности диктора и измеряемых по одной или нескольким фразам. Измеренные значения сравниваются (часто с использованием некоторых существенно нелинейных



метрик близости) с аналогичными параметрами эталонных образов подлежащего опознанию диктора.

Таким образом, при верификации диктора требуется однократное сравнение совокупности (совокупностей) измеренных значений со значениями параметров эталонов, на основе которого выносится решение о принятии или отклонении.

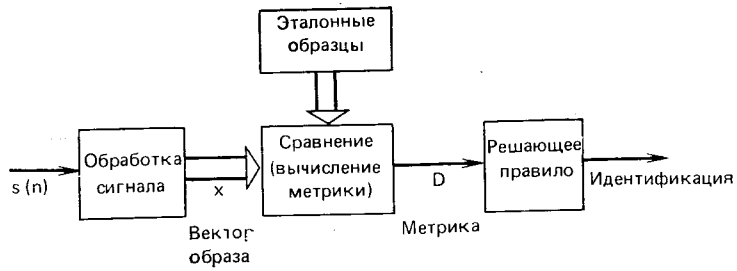


Рис. 9.10. Общее представление задачи распознавания диктора

предполагаемой идентичности. В общем случае вычисляется расстояние между измеренными значениями и распределением эталонов. На основе распределения потерь между возможными типами ошибок (т. е. верификации «самозванца» и отклонении «подлинного» диктора) устанавливается соответствующий порог различимости (расстояния). Вероятность перечисленных выше ошибок практически не зависит от $N$ (числа эталонов, хранимых в системе), поскольку все эталоны голосов других дикторов используются для формирования устойчивого распределения, характеризующего всех дикторов. Записывая сказанное выше в математической форме, обозначим распределение вероятности измеренных значений вектора $\mathbf{x}$ для диктора как $p_i(\mathbf{x})$, что приводит к простому решающему правилу вида

$$\begin{aligned} &\text{Верифицировать диктора } i, \text{ если } p_i(\mathbf{x}) > c_i\, p_{av}(\mathbf{x}); \\ &\text{Отклонить диктора } i, \text{ если } p_i(\mathbf{x}) < c_i\, p_{av}(\mathbf{x}), \end{aligned} \tag{9.1}$$

где $c_i$ — константа для $i$-го диктора, определяющая вероятности ошибок $i$-го диктора, а $p_{av}(\mathbf{x})$ — среднее (по всему ансамблю дикторов) распределение вероятности измеренных значений вектора $\mathbf{x}$. Изменяя порог $c_i$, можно изменять вероятность ошибки, определяемую вероятностями ошибок обоих типов.

Задача идентификации диктора существенно отличается от задачи верификации. В этом случае система должна точно указать одного из дикторов среди $N$ дикторов данного множества. Таким образом, вместо однократного сравнения измеряемых параметров с хранимым в системе эталоном необходимо провести $N$ сравнений. Решающее правило в этом случае сводится к *выбору такого диктора i, для которого*

$$\begin{aligned} &p_i(\mathbf{x}) > p_j(\mathbf{x}), \\ &j = 1,\ 2,\ \ldots.\ N,\ j \neq i, \end{aligned} \tag{9.2}$$

т. е. выбирается диктор с минимальной абсолютной вероятностью ошибки. С увеличением количества дикторов в ансамбле возрастает и вероятность ошибки, поскольку большое число вероятностных распределений в ограниченном пространстве параметров не может не пересекаться. Все более вероятным становится то, что два или более дикторов в общем ансамбле будут иметь распределения вероятностей, которые близки друг к другу. При таких условиях приемлемая идентификация дикторов становится практически невозможной.

Приведенный выше анализ позволяет сделать вывод, что между задачами идентификации и верификации имеется много общего и много различий. В каждом случае диктор должен произнести одну или несколько тестовых фраз. По этим фразам проводятся некоторые измерения, и затем вычисляются одна или

несколько мер различимости («расстояния») между предъявленным и эталонным векторами. Таким образом, с позиции методов цифровой обработки обе эти задачи сходны. Основное различие возникает на этапе вынесения решений.

### 9.2.1. Система верификации диктора

На рис. 9.11 показана структурная схема системы верификации диктора в реальном масштабе времени [13—16]. Лицо, желающее быть верифицированным, сначала вводит в систему данные, подтверждающие его право на идентификацию, а затем по

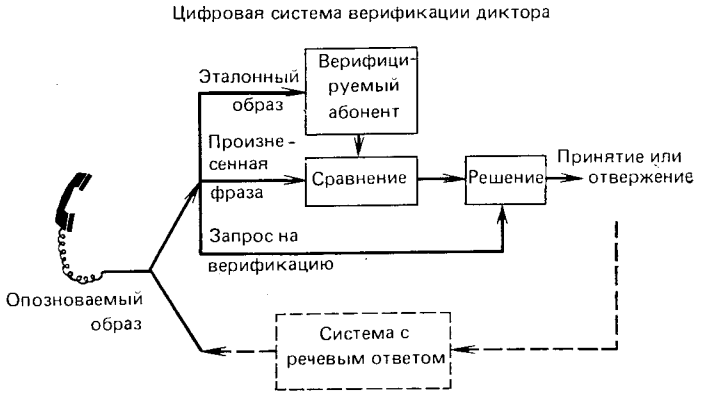


Рис. 9.11. Структурная схема системы верификации диктора [13]

запросу системы (сформированному, например, системой с речевым ответом) произносит эталонные фразы и в случае его верификации поручает системе выполнить необходимые операции. Система обрабатывает произнесенную диктором тестовую фразу с целью получения образа, сравниваемого затем с эталонным образом, соответствующим указанному при передаче права на идентификацию. Затем вычисляется матрица потерь [$c_i$ в (9.1)], определяющая полную ошибку, и выносится решение о принятии или отклонении утверждения абонента об идентичности.

На рис. 9.12 представлена та часть системы верификации, в которой непосредственно осуществляется обработка сигнала. Отсчеты речевого сигнала, возникающие где-либо внутри выбранного интервала, обрабатываются с целью определения начала и конца фразы. Это осуществляется в устройстве анализа моментов начала и конца фразы. Такое устройство описано в гл. 4. После определения начала и конца фразы проводится ряд измерений и оценок параметров для формирования образа, описывающего данную фразу. Обычно используются измерения следующих параметров: периода основного тона для получения траектории периода основного тона данной фразы; кратковременной энергии для получения траектории кратковременной энергии; коэффициентов линейного предсказания для получения траектории передаточной



функции речеобразующего тракта и, наконец, оценивание формантных частот. (Все показанные на рис. 9.12 параметры одновременно используются в системе верификации, однако вследствие

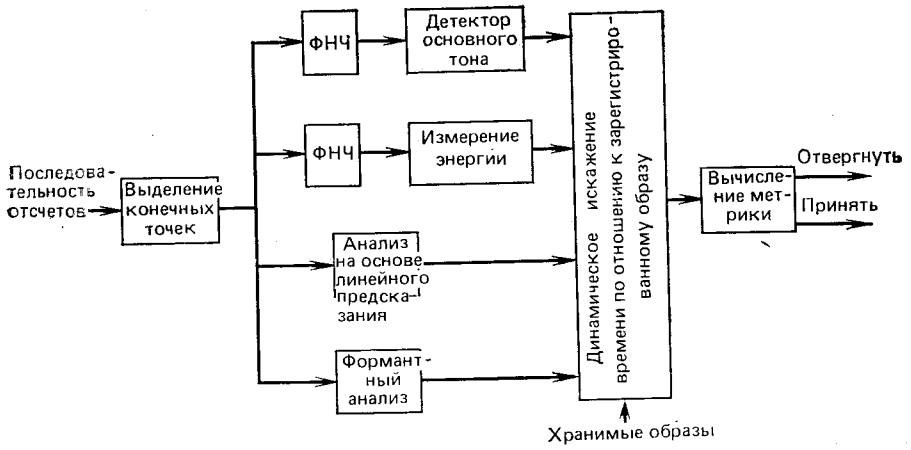


Рис. 9.12. Основные этапы обработки речевого сигнала в системе верификации диктора

большого объема вычислений, которые необходимо осуществить для анализа по методу линейного предсказания или для оценивания формантных частот при верификации в реальном масштабе

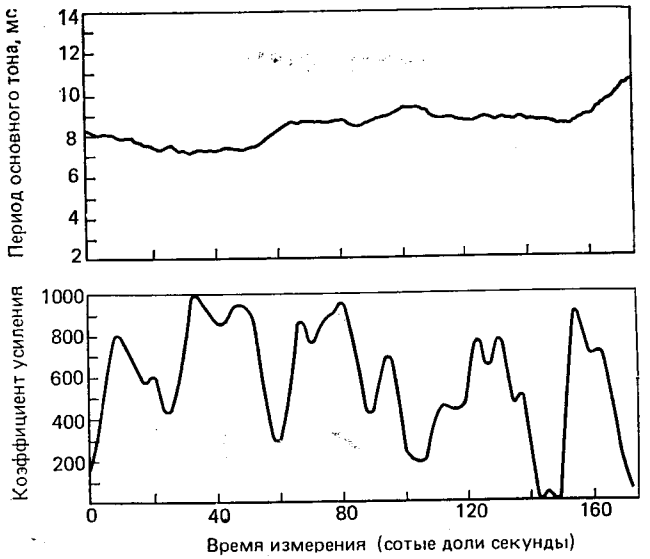


Рис. 9.13. Период основного тона и энергетический контур, используемые при верификации диктора [13]

времени, ограничиваются только измерением основного тона и энергии.)

Для иллюстрации широких возможностей использования методов цифровой обработки в таких системах полезно упомянуть конкретные алгоритмы, использованные Розенбергом (и другими разработчиками этой системы) при проведении соответствующих измерений. Выделение основного тона осуществлялось на основе методов параллельной обработки во временно́й области (см. гл. 4). Интенсивность измерялась с использованием кратковременного усреднения абсолютного значения в соответствии с результатами гл. 4. Для анализа по методу линейного предсказания использовался автокорреляционный алгоритм, рассмотренный в гл. 8. Наконец, формантный анализ проведен на основе гомоморфной фильтрации, описанной в гл. 7.

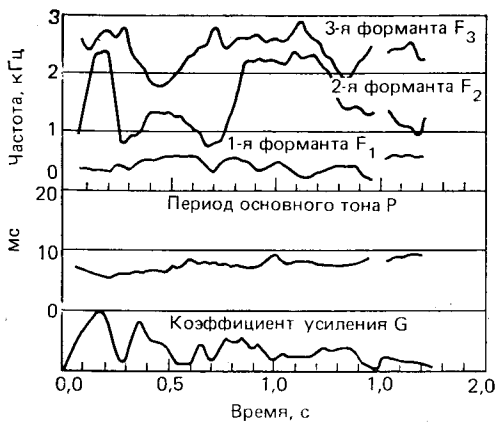


Рис. 9.14. Траектории первых трех формант, период основного тона, интенсивность для верификации диктора [15]

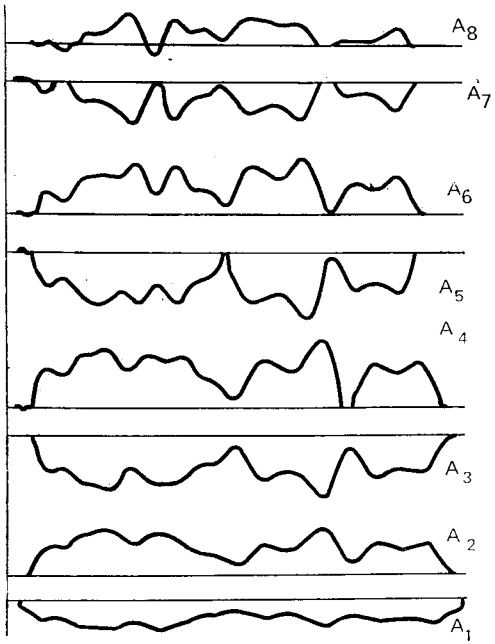


Рис. 9.15. Траектории первых восьми коэффициентов линейного предсказания для верификации диктора по фразе [16]

На рис. 9.13—9.15 показаны типичные траектории измерений для тестовой фразы: «We were away a year ago», произнесенной мужским голосом. На рис. 9.13 представлены траектории периода основного тона и интенсивности по всей фразе [13]. Эти данные оценивались периодически 100 раз/с и сглаживались фильтром нижних частот КИХ-типа с полосой 16 Гц. Для данного диктора флуктуации траектории интенсивности значительно превосходят флуктуации в траектории основного тона. На рис. 9.14 для той же фразы, произнесенной другим диктором, представлены траектории трех первых формант совместно с контуром основного тона и интенсивности [15]. Формантные траектории сглажены таким же КИХ-фильтром с частотой среза 16 Гц. Наконец, на рис. 9.15 представлены первые восемь коэффициентов предсказания 12-полюсной модели [16]. Из этого рисунка видно, что для данной фразы описание с помощью параметров линейного предсказания обладает значительной избыточностью. Таким образом, при использовании этих данных с целью верификации можно утверждать, что коэффициенты линейного предсказания внесут меньший вклад в уменьшение ошибок верификации. Можно считать поэтому, что при правильном подборе оцениваемых параметров при верификации можно получить почти такую же вероятность ошибки, как и при совместном использовании всех оценок, перечисленных выше.

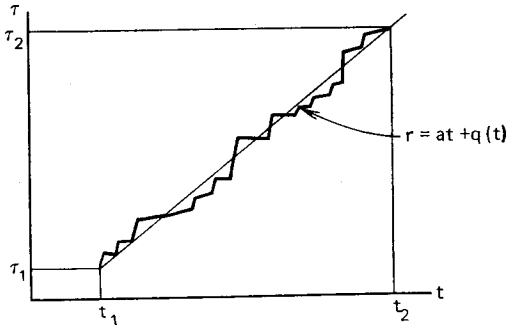


Рис. 9.16. Функция преобразования временно́го масштаба

После вычисления необходимых оценок параметров их необходимо сравнить с соответствующими эталонами голоса идентифицируемого диктора. Поскольку диктор не в состоянии повторить абсолютно точно в том же темпе одну и ту же фразу, нецелесообразно сравнивать такие временны́е параметры, как период основного тона, интенсивность и изменение во времени формантных частот. Эту трудность можно преодолеть путем нелинейного преобразования временно́го масштаба входного множества параметров для получения более точного соответствия между эталоном и последующими оценками параметров для одного и того же диктора. Процесс преобразования временно́го масштаба чрезвычайно важен и часто используется при обработке речевого сигнала.

Процесс преобразования временно́го масштаба схематически представлен на рис. 9.16. Временно́й масштаб следует трансформировать таким образом, чтобы характерные точки измеренной траектории $a(t)$ совпали с характерными точками эталонной траектории $r(t)$. Предполагается, что преобразующая функция имеет вид

$$\tau = \alpha t + q(t), \tag{9.3}$$

где $q(t)$ — нелинейная функция трансформации масштаба, а $\alpha$ представляет собой средний наклон характеристики преобразования. Отсутствие $q(t)$ соответствует простой линейной модификации. Граничные условия накладываются таким образом, чтобы

начальная и конечная точки исходной и эталонной фраз точно совпали. Эти условия имеют вид

$$\tau_1 = \alpha t_1 + q \mid t_1); \tau_2 = \alpha t_2 + q \mid t_2). \qquad (9.4а, б)$$

Теперь осталось только выбрать константу и параметр $q(t)$ таким образом, чтобы достичь наилучшего совпадения сравниваемых траекторий. Один из наиболее простых подходов заключается в определении $q(t)$ как кусочно-линейной функции с конечным числом точек излома по оси $t$, в которых изменяется наклон $q(t)$. Точки излома и наклон $q(t)$ (как и средний наклон $\alpha$) определяются затем методом наискорейшего спуска, где в качестве критерия выступает мера различимости либо в виде расстояния между обрабатываемой и эталонной траекториями, либо в виде корреляции между ними.

Значительно более простым и эффективным с точки зрения вычислений является метод динамического программирования для оптимального выбора функции, преобразующей временной масштаб. Наложив вместо условия кусочной линейности условие непрерывности, относительно несложно определить оптимальный алгоритм преобразования для множества траекторий [17].

Рассмотрим работу алгоритма преобразования масштаба для двух траекторий, представленных в виде дискретной последовательности отсчетов. Обозначим точки на измеренной траектории через $n=1, 2, ..., N$, а точки на эталонной траектории через $m=1, 2, ..., M$. Требуется выбрать такую функцию преобразования масштаба $w$, чтобы выполнялись условия:

$$\left.\begin{array}{l} m = w(n);\ w(1) = 1 \text{ в начальной точке,} \\ w(N) = M \text{ в конечной точке,} \end{array}\right\} \qquad (9.5),\ (9.6)$$

Если используемая преобразующая функция линейна, то она имеет вид

$$w(n) = \left[\left(\frac{M-1}{N-1}\right)(n-1) + 1\right]. \qquad (9.7)$$

Если для преобразования используется нелинейная функция, то в соответствии с граничными условиями следует рассмотреть стратегию движения из начальной точки $n=1$, $m=1$ в конечную точку $n=N$, $m=M$ по дискретной сетке точек на плоскости. Для ограничения степени нелинейности преобразующей функции целесообразно предположить, что $w$ не может изменяться более чем на два шага дискретной сетки при любом $n$. Иными словами,

$$w(n+1) - w(n) = \begin{cases} 0, 1, 2, & w(n) \neq w(n-1); \\ 1, 2, & w(n) = w(n-1). \end{cases} \qquad (9.8)$$

Таким образом, при изменении значения функции в предшествующей точке ее приращения в данной точке могут составлять 0, 1, 2, а в противном случае — только 1 и 2. Чтобы определить, какие из условий (9.8) выполняются, необходимо иметь меру сходства между эталонной траекторией в точке $n$ и входной траекторией в точке $m$. Мера сходства (или расстояние) между траекториями используется для определения вида преобразующей функции, которая доставляет локальный минимум максимальному значению расстояния по всей траектории в соответствии с ограничениями (9.8).



Для примера на рис. 9.17 [17] показаны область возможных значений дискретной сетки $(n, m)$ и типичная функция преобразования масштаба $w(n)$ (сплошная линия внутри сетки) для трансформации 20 точек ($N=20$) эталонной траектории в 15 точек ($M=15$) обрабатываемой траектории. Вследствие ограничений непрерывности получающаяся функция должна лежать внутри параллелограмма, изображенного на рисунке.

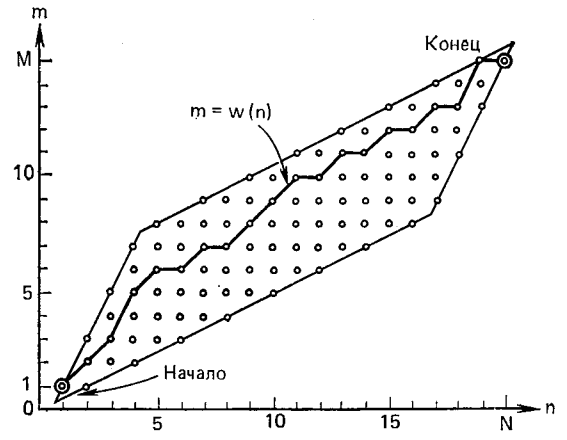


Рис. 9.17. Пример построения функции преобразования временно́го масштаба [17]

Метод трансформации временно́го масштаба был использован как для верификации [13], так и для распознавания речи [17]. В качестве примера на рис. 9.18 представлены траектории интенсивности как до, так и после трансформации масштаба [13]. В данном случае сближение траекторий весьма заметно.

Заключительным шагом в процессе верификации (см. рис. 9.12) являются вычисление некоторой полной меры различимости (на основе частных мер различимости отдельных траекторий) и сравнение ее с выбранным соответствующим образом порогом. Простейшей мерой различимости двух траекторий может служить нормированная сумма квадратов; например, для $j$-й траектории мера различимости $d_j$ будет иметь вид

$$d_j = \sum_i [[a_{js}(i) - a_{jr}(i)]/\sigma_{aj}(i)]^2, \qquad (9.9)$$

где $a_{js}(i)$ — значение $j$-й траектории входного сигнала в момент $i$; $a_{jr}(i)$ — значение $j$-й траектории эталона в момент $i$; $\sigma_{aj}(i)$ —

стандартное отклонение $j$-й траектории в момент $i$. Полная мера различимости обычно представляет собой взвешенную сумму корней, т. е.

$$D = \sum_i w_j d_j, \tag{9.10}$$

где $w_j$ — вес, выбираемый на основе значимости $j$-го измеренного значения траектории верифицируемого диктора.

Рассмотренная выше система верификации диктора была всесторонне исследована, и полученные результаты позволили сделать вывод о том, что они являются потенциально достижимыми для такого рода систем. Был проведен целый ряд экспериментов по проверке системы как при использовании высококачественных фраз при малом числе дикторов, так и при использовании фраз «телефонного» качества при очень большом числе дикторов. Использовался даже хорошо тренированный имитатор (подражатель), пытавшийся «обмануть» систему. Результаты этих экспериментов показали, что в случае высококачественного сигнала равные вероятности ошибок в системе (т. е. когда вероятность ошибочного отклонения диктора равна вероятности ошибочного отождествления) могут быть сделаны сколь угодно близкими к нулю, если выбрано достаточное количество наблюдений и подобраны весовые коэффициенты для каждого диктора. Равновероятные ошибки в таких системах при использовании профессионального имитатора составляют 4,1%. При использовании сигнала телефонного качества и применении в качестве признаков интенсивности и периода основного тона вероятность ошибок составила примерно 7%. Добавление более сложных признаков, таких, как форманты или параметры линейного предсказания, приводит к значительному уменьшению вероятности ошибки.

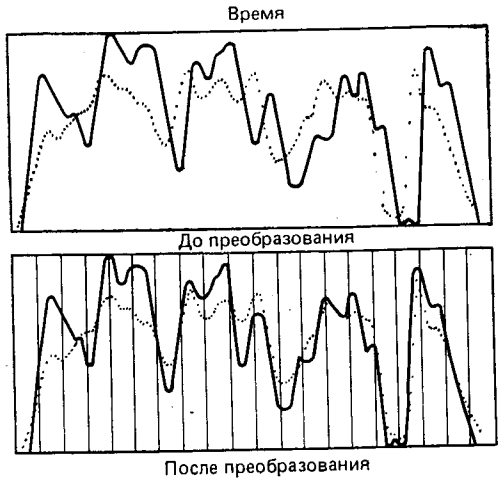


Рис. 9.18. Пример преобразования временно́го масштаба траектории кратковременной энергии речи [13]

### 9.2.2. Система идентификации диктора

Задачи идентификации и верификации во многом сходны. С точки зрения обработки сигналов обе рассматриваемые задачи почти совпадают и практически все, что изображено на рис. 9.12,



в равной мере подходит как для верификации, так и для идентификации. Основное отличие заключается в тех параметрах, которые используются для построения меры различимости, а также в необходимости вычисления $N$ мер различимости вместо одной. Решение, формируемое системой при идентификации диктора, сводится к выбору того диктора, чье эталонное описание наиболее близко к описанию, полученному по входному сигналу. При верификации требуется решить задачу бинарного выбора, т. е. принять или отклонить утверждение о том, что голос опознаваемого диктора идентичен данному эталону, и это достигается на основе сравнения значения меры различимости с выбранным порогом.

Хотя для целей верификации систем вполне пригодна классическая мера различимости (9.10), для идентификации диктора обычно используют более сложные и устойчивые к различным аномалиям меры различимости [18, 20]. Напомним, что значение меры различимости вычисляется с целью сравнения входного и эталонного образов. Мера различимости, используемая Аталом [18, 10], может быть получена следующим образом. Пусть $\mathbf{x}$ представляет собой вектор-столбец измеренных значений входного сигнала размерностью $L$, причем элементом $\mathbf{x}$ является $k$-е измеренное значение. Предполагается, что совместная функция плотности вероятности измеренных значений для $i$-го диктора представляет собой многомерное распределение Гаусса со средним значением $m_i$ и ковариационной матрицей $\mathbf{W}_i$. Таким образом, $L$-мерная плотность распределения Гаусса для $\mathbf{x}$ имеет вид

$$g_i(\mathbf{x}) = (2\pi)^{-1/2} |\mathbf{W}_i|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-m_i)^t W_i^{-1}(\mathbf{x}-m_i)\right], \tag{9.11}$$

где $\mathbf{W}_i^{-1}$ — матрица, обратная $\mathbf{W}_i$ ($\mathbf{W}_i$ предполагается неособенной); $|\mathbf{W}_i|$ — детерминант $\mathbf{W}_i$; $t$ — транспонирование вектора. Решающее правило, минимизирующее вероятность ошибки, состоит в том, что вектор измеренных значений $\mathbf{x}$ следует отнести к классу $i$, если

$$p_i g_i(\mathbf{x}) \geqslant p_j g_j(\mathbf{x}), \; i \neq j, \tag{9.12}$$

где $p_i$ — априорная вероятность принадлежности вектора $\mathbf{x}$ к классу $i$. Поскольку $\ln y$ — монотонно возрастающая функция своего аргумента, решающее правило (9.12) можно значительно упростить, переписав в виде

$$d_i(\mathbf{x}) = \begin{cases} \frac{1}{2}(\mathbf{x}-m_i)^t \mathbf{W}_i^{-1}(\mathbf{x}-m_i) + \frac{1}{2}\ln|W_i| = \ln p_i \leqslant \\ \leqslant d_j(\mathbf{x}), \quad i \neq j. \end{cases} \tag{9.13}$$

Последние два члена в правой части (9.13) не зависят от вектора $\mathbf{x}$, и поэтому можно считать, что они представляют собой смещение $i$-го класса. Для большинства практически важных случаев установлено, что решающее правило со смещением в правой части не имеет преимуществ перед решающим правилом, основанным

только на первом члене (9.13). Таким образом, функцию различимости можно определить как

$$\hat{d}_i = (\mathbf{x} - m_i)^t \mathbf{W}_i^{-1} (\mathbf{x} - m_i), \qquad (9.14)$$

а индекс $i$ выбран таким образом, чтобы минимизировать по этому индексу.

Решающее правило предполагает вычисление вектора средних и ковариационной матрицы для каждого класса $i$ на множестве решения. Вектор средних и ковариационная матрица определяются по обучающей последовательности $\mathbf{x}_i(n)$ векторов, принадлежащих $i$-му классу:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathbf{x}_i(n) \qquad (9.15)$$

и

$$\mathbf{W}_i = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathbf{x}_i(n)\, \mathbf{x}_i^t(n) - \mathbf{m}_i\, \mathbf{m}_i^t. \qquad (9.16)$$

На рис. 9.19 представлены типичные примеры распределений параметров, измеренных по речевому сигналу, и их приближений одномерными гауссовскими распределениями. В одних случаях степень согласия больше, чем в других.

Следует сказать несколько слов относительно предположений и необходимых вычислений, приводящих к решающему правилу (9.14). Предположение относительно нормальности распределения измеренных значений можно подтвердить рядом соображений. Во-первых, чтобы решающее правило оставалось в силе, распределение не обязательно должно быть строго нормальным. Эта особенность часто проявляется в физических измерениях. Например, в случае унимодальных распределений достаточно, чтобы распределение было нормально в центральной области возможных значений. Более того, как отмечалось выше, решающее правило оказывается оптимальным для целого класса распределений, которые могут быть получены из нормального с помощью монотонных преобразований. Наконец, решающее правило требует знания только первых двух моментов распределения. Точное оценивание высших моментов оказывается задачей трудноразрешимой на практике.

Важное преимущество функции различимости (9.14) состоит в ее инвариантности к несингулярным линейным преобразованиям [20]. Это свойство инвариантности оказывается чрезвычайно важным, поскольку использование некоторой совокупности параметров и их линейного преобразования приводит к одним и тем же результатам, например одинаковые результаты можно получить по некоторым траекториям и их преобразованию Фурье. Второе важное свойство меры различимости (9.14) состоит в том, что она построена на основе взвешивания различных компонент опознаваемого вектора в соответствии с их значимостью [20].



Используя решающее правило (9.14), Атал исследовал эффективность различных параметрических представлений речевого сигнала применительно к задаче идентификации диктора [20]. Каждый из десяти дикторов произносил один и тот же текст по 6 раз.

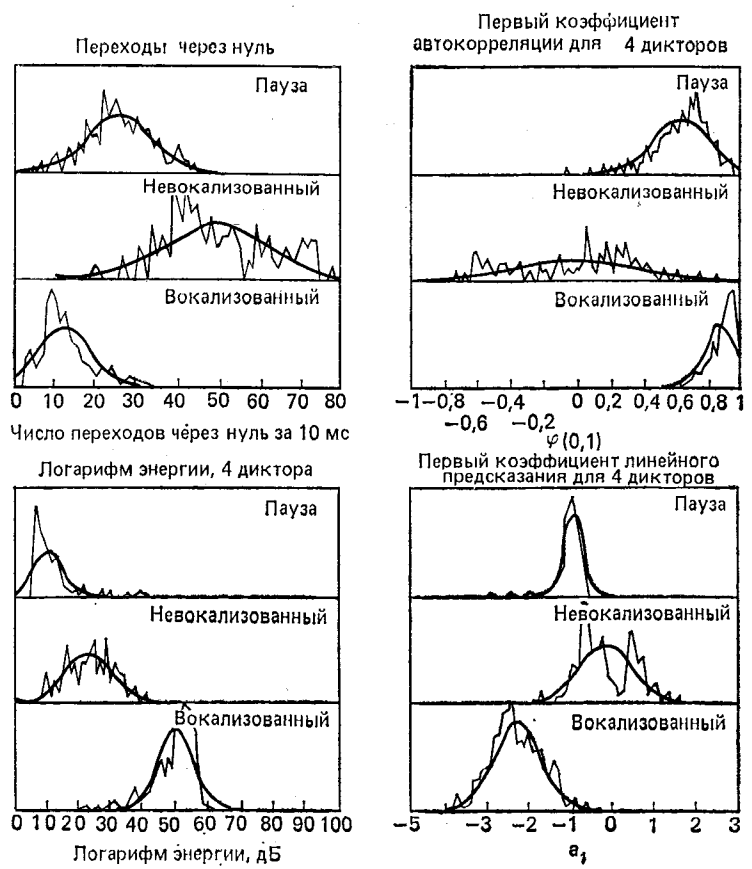


Рис. 9.19. Измеренные распределения некоторых параметров речевого сигнала с подогнанными к ним гауссовыми кривыми [27]

Каждая фраза разделялась на 40 сегментов равной длины, чем обеспечивалось грубое выравнивание масштаба времени. Средняя длина каждого сегмента составляла около 50 мс. Затем проводился анализ на основе линейного предсказания на каждом из 40 сегментов для каждого из 60 предложений. Таким образом был получен вектор образов для каждого интервала анализа. По коэффициентам линейного предсказания рассчитывались импульсная характеристика, автокорреляционная функция, функция площадей поперечного сечения в неоднородной акустической трубе без потерь и кепстр. Затем проверялась точность идентификации, для чего одна фраза служила опознаваемой, а остальные использова-

лись как эталонные для каждого диктора. Решающее правило (9.14) использовалось для идентификации диктора на каждом из 40 интервалов анализа с целью определения вероятности правильной идентификации. Результаты для каждого из параметров представлены на рис. 9.20.

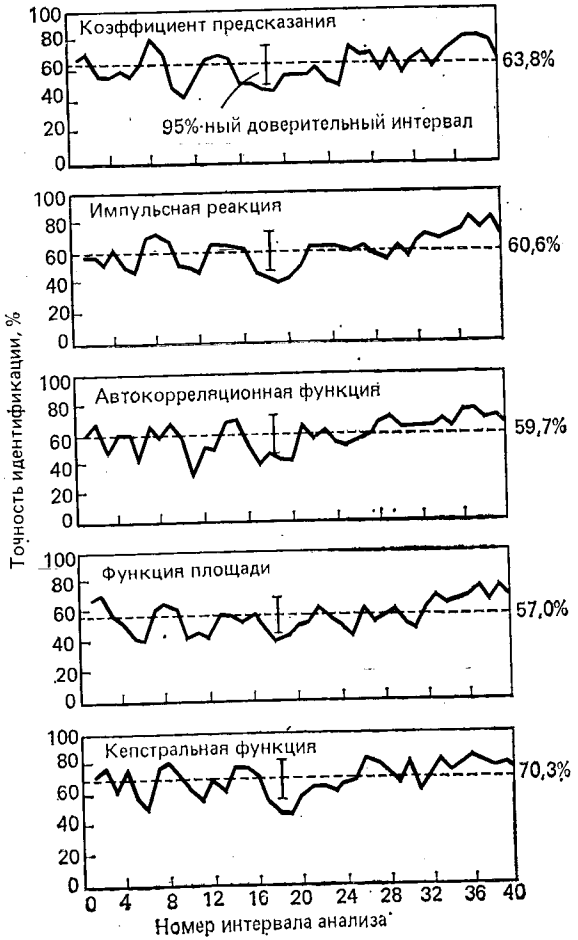


Рис. 9.20. Точность идентификации диктора в зависимости от множества используемых параметров [20]

Все параметры дают примерно одинаковую вероятность ошибки, однако точность кепстрального метода несколько выше, чем всех других. Объединяя несколько интервалов анализа для получения описания опознаваемого вектора большей размерности, можно добиться меньшей вероятности ошибки. На рис. 9.21 представлены кривые точности идентификации, достигнутой с использованием кепстрального метода в зависимости от длительности



сегмента речевого сигнала, используемого для вычисления различимости. Полученные результаты показывают, что для данного ансамбля дикторов 95%-ная точность идентификации может быть достигнута на сегментах сигнала длительностью около 0,5 с.

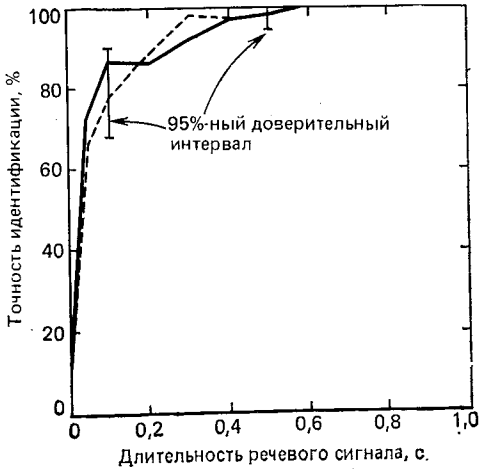


Рис. 9.21. Точность идентификации диктора (с использованием кепстральных параметров) в зависимости от длительности речевого сигнала [20]

## 9.3. Системы распознавания речи [17, 21—26]

Как и при распознавании диктора, методы цифровой обработки применяются при распознавании речевого сигнала для получения описания распознаваемого образа, которое затем сравнивается с хранимыми в памяти эталонами. Задача распознавания речевого сигнала состоит в определении того, какое слово, фраза или предложение были произнесены.

В отличие от областей машинного речевого ответа и распознавания диктора, где задача в общем случае достаточно определена, область распознавания слов является одной из тех, где, прежде чем поставить задачу, требуется ввести большое число предположений, например:

тип речевого сигнала (изолированные слова, непрерывная речь и т. д.);

число дикторов (система для одного диктора, нескольких дикторов, неограниченного числа дикторов);

тип диктора (определенный, случайный, мужчина, женщина, ребенок);

условия произнесения фраз (звукоизолированное помещение, машинный зал, общественное место);

система передачи (высококачественный микрофон, узконаправленный микрофон, телефон);

тип и число циклов обучения (без обучения, с ограниченным числом циклов обучения, с неограниченным числом циклов обучения);

размер словаря (малый объем 80—20 слов, средний объем 20—100 слов и большой объем — более 100 слов).

формат произносимых фраз (ограниченный по длительности текст, свободный речевой формат).

Из приведенного перечня условий следует, что при создании систем распознавания речи реализация некоторых из условий может оказаться более предпочтительной. В данном параграфе будут рассмотрены три наиболее распространенных типа систем распознавания речи, в которых широко используются методы цифровой обработки сигналов. Все они являются системами распознавания с ограниченным словарем, не содержащим контекста. Хотя в системах распознавания слитной речи также широко используются цифровые методы обработки [21, 22], однако большая часть усилий при разработке таких систем затрачивается на синтаксический и семантический анализ фраз. Эти области близко примыкают к лингвистической теории речи, поэтому изложение подобных вопросов увело бы нас в сторону от рассматриваемых здесь задач. Интересующегося читателя мы отсылаем к соответствующей литературе, содержащей обсуждение систем, «понимающих» речь.

### 9.3.1. Система распознавания изолированных цифр [25]

Система распознавания изолированных цифр обладает следующими свойствами:

1. Словарь малого объема состоит из изолированных слов, обозначающих десять цифр (0—9).
2. Отсутствуют ограничения на количество дикторов, а также на их пол и возраст.
3. Условия произнесения: машинный зал, микрофон — узконаправленный или высококачественный.
4. Обучение не предусмотрено.
5. Формат на входе — однословный с паузами между словами.

На рис. 9.22 представлена структурная схема системы распознавания изолированных цифр. Как видно из этого рисунка, основными элементами системы являются устройство анализа моментов

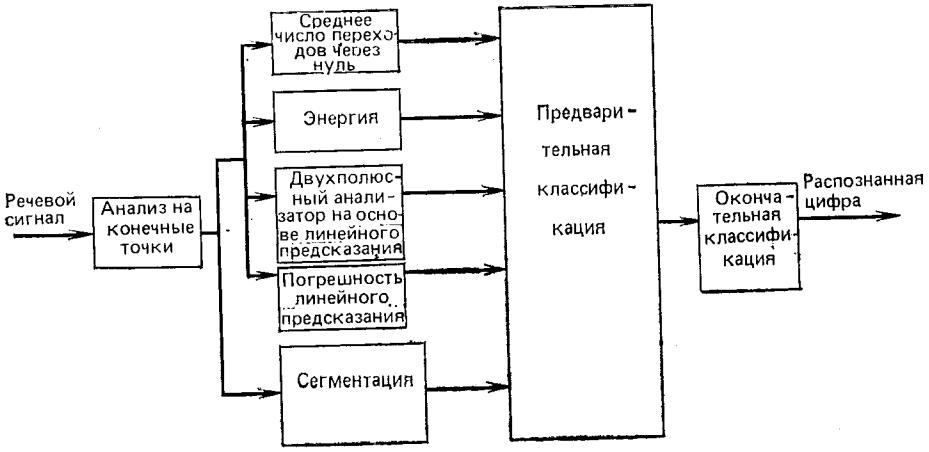


Рис. 9.22. Структурная схема системы распознавания отдельно произнесенных цифр [25]

начала и окончания слова (как и в системе распознавания), устройство обработки, формирующее образ или вектор измеренных значений, устройство сегментации фразы на интервалы и блок предварительных и окончательных решений относительно произнесенной цифры.

Хотя существует много способов представления сигнала, которые можно использовать в системах распознавания речи, представления, применяемые в системах, инвариантных к диктору, должны быть достаточно устойчивыми [25]. Измерения параметров должны быть простыми и однозначными, а их измеренные значения должны наиболее полно отражать различия в звуках речи.

Кроме того, измерения должны допускать достаточно простую интерпретацию с позиций систем, инвариантных к диктору. В одной из таких устойчивых систем (см. рис. 9.22) использованы следующие параметры: среднее число переходов через нуль, энергия, коэффициенты линейного предсказания с использованием двухполюсной модели и погрешность предсказания.



Измерения первых двух параметров рассматривались в гл. 4. Хотя в гл. 8 рассматривался общий метод линейного предсказания, использование двухполюсной модели является несколько необычным. Использование двухполюсной модели было предложено Макоулом и Вульфом [26] как удачный метод описания основных свойств кратковременного спектра. Частота полюса характеризует основную концентрацию энергии в спектре, а погрешность предсказания показывает общий наклон спектра.

Для иллюстрации сказанного на рис. 9.23 [26] представлены спектры отдельных звуков речи, полученные с использованием

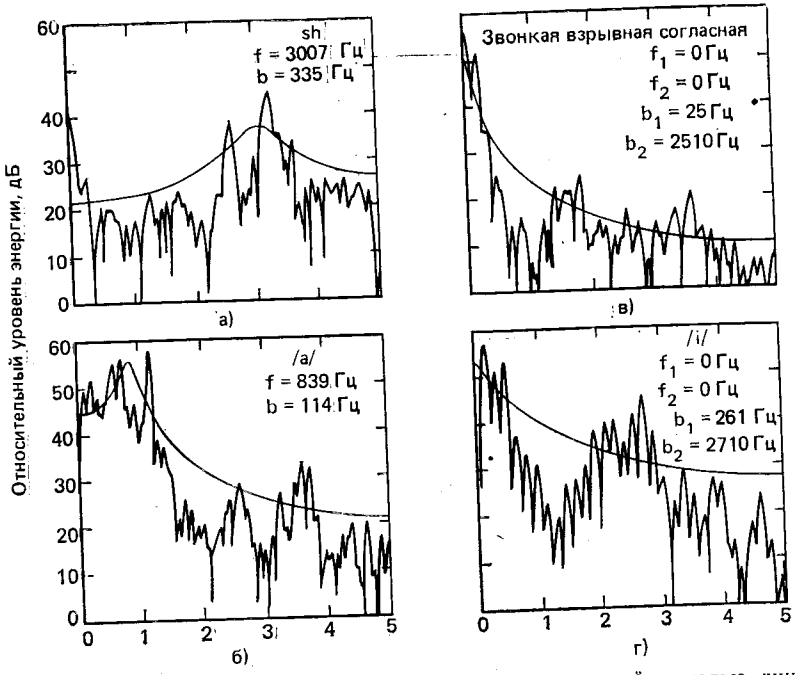


Рис. 9.23. Согласие между спектром сигнала и двухполюсной моделью линейного предсказания для некоторых звуков речи [26]

БПФ, и спектры соответствующей этим звукам двухполюсной модели. Для случая двухполюсного анализа полином имеет либо один комплексно-сопряженный корень, либо два действительных корня. На рис. 9.23*а* изображен спектр звука */sh/* в слове «short». В данном примере двухполюсная модель дает комплексно-сопряженный корень на частоте около 3 кГц, т. е. в области максимальной концентрации энергии в спектре. На 9.23*б* представлены аналогичные результаты для гласного звука */a/*, где концентрация

основной энергии в спектре имеет место на частоте около 800 Гц. В примерах рис. 9.23*в* основная часть энергии спектра сосредоточена в области нулевых частот и, таким образом, модель имеет два действительных полюса в правой полуплоскости $z$-плоскости.

Из рис. 9.23 видно, что вычисленные частоты двухполюсной модели хорошо описывают распределение энергии в спектре звука и могут, таким образом, быть использованы для описания звуков с относительно высоко- или низкочастотным распределением энергии. Так, например, шумоподобные звуки характеризуются относительно высокочастотной концентрацией энергии, тогда как носовые и гласные звуки имеют относительную концентрацию в области низких частот.

Для иллюстрации типичных результатов анализа на рис. 9.24 и 9.25 представлены траектории параметров, построенных по словам «девять» и «шесть». Предварительное распознавание основано

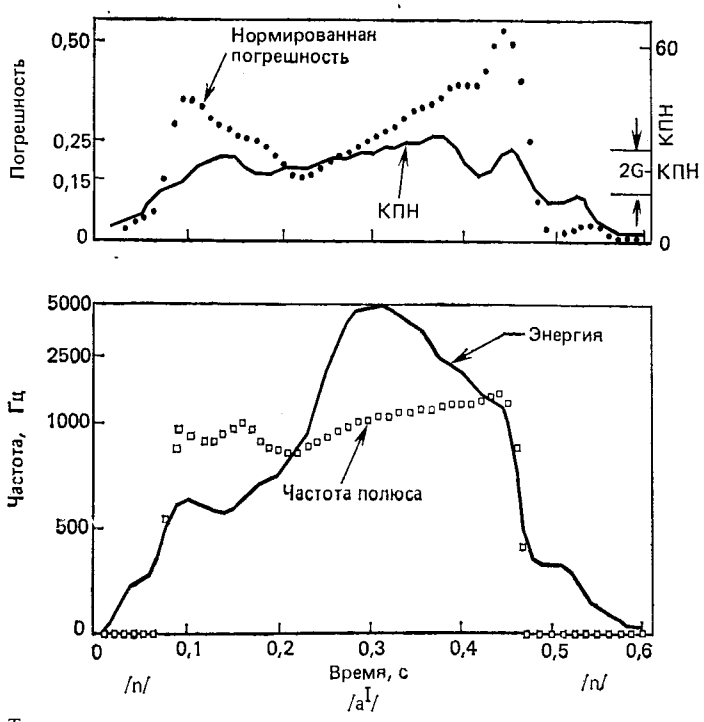


Рис. 9.24. Траектории сигналов для распознавания отдельно произнесенной цифры /nine/ [25]

на грубой классификации цифр на основе анализа каждого отдельного измерения (в различных точках на протяжении всего слова), а окончательное решение выносится путем объединения отдельных решений по каждому из измеренных значений. Так, например, начальный носовой сегмент слова «девять» на рис. 9.24



характеризуется малой нормированной погрешностью предсказания и положением полюсов на нулевой частоте, в то время как фрикативным началу и концу слова «шесть» на рис. 9.25 соответствуют значительная нормированная погрешность предсказания, отличная от нуля частота полюсов спектра и большое количество переходов через нуль (КПН).

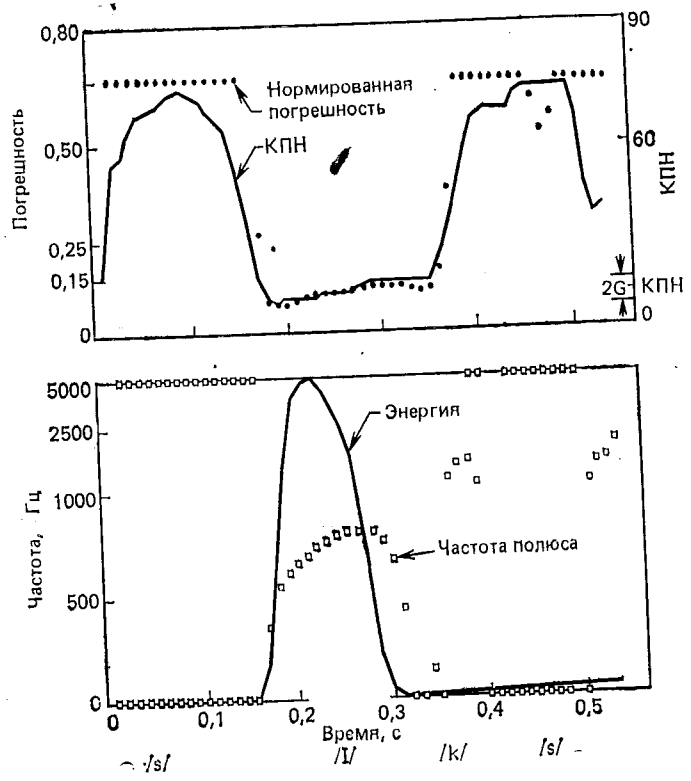


Рис. 9.25. Траектории сигналов для распознавания отдельно произнесенной цифры /six/ [25]

Самбуром и Рабинером [25] предложен древовидный алгоритм принятия окончательного решения на основе совместной обработки частных решений, полученных по каждому измеренному значению на каждом интервале анализа. При использовании этого алгоритма точность распознавания для 65 дикторов составила от 94,4 до 97,3%.

### 9.3.2. Система распознавания слитной последовательности цифр

Приведем решение более сложной задачи распознавания слитной последовательности цифр при произнесении их произвольным диктором. Свойства, которыми должна обладать эта система, в

основном совпадают со свойствами системы распознавания, рассмотренной в 9.3.1, с одним важным исключением. Свойство 5 в данном случае состоит в необходимости распознавания слитной последовательности из трех слов (цифр) без пауз между ними.

Хотя между системами распознавания изолированных цифр и слитной последовательности цифр много общего, реализация этих систем распознавания существенно различается, особенно в блоке анализа или обработки сигнала. Это связано с необходимостью предварительной сегментации слитной последовательности на отдельные цифры перед их распознаванием. Задача сегментации является чрезвычайно сложной, и в настоящее время не найдено простого решения для общего случая.

На рис. 9.26 представлена структурная схема блока обработки сигнала в системе распознавания слитной последовательности

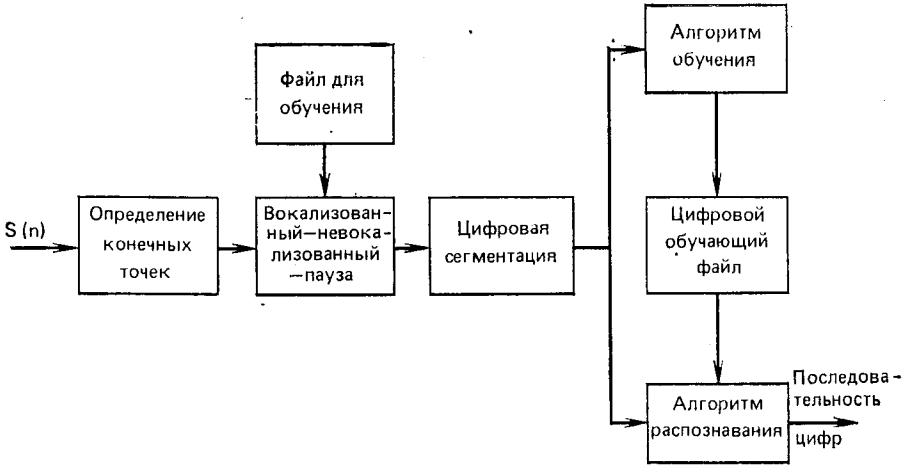


Рис. 9.26. Структурная схема системы распознавания последовательности цифр [27]

цифр. Записанная последовательность цифр первоначально подвергается анализу с целью определения моментов начала и окончания фразы. Для этого используется метод анализа во временной области, изложенный в гл. 4. Вслед за определением моментов начала и окончания фразы речевой сигнал подвергается обработке с целью оценивания следующих параметров (100 раз/с): среднего числа переходов через нуль, логарифма энергии, коэффициентов линейного предсказания, логарифма погрешности линейного предсказания и первого коэффициента автокорреляции.

Измеренные параметры используются затем в качестве входных сигналов решающего устройства, которое классифицирует каждый 10-миллисекундный интервал как вокализованный, невокализованный или паузу на основе неевклидовой метрики [такого типа, как (9.14)]. Для сегментации слитного потока на отдельные



цифры используется нелинейно сглаженная траектория признака «вокализованный—невокализованный—пауза» совместно с некоторой статистической информацией о надежности классификации на каждом интервале и измеренными значениями энергии сигнала. Для правильной сегментации в систему необходимо ввести информацию о количестве цифр в фразе. Для всех примеров, рассматриваемых в данном разделе, предполагается, что последовательность содержит ровно три цифры.

Сегментация осуществляется на основе использования известных результатов по различным измерениям на каждом 10-миллисекундном интервале. Например, известно, что невокализованный интервал соответствует интервалу, в пределах которого находится искомая граница, поскольку ни одна из цифр не содержит невокализованных звуков внутри слова. Известно также, что глубокие провалы в траектории энергии на вокализованном сегменте почти всегда соответствуют границе между цифрами. Основываясь на этих наблюдениях, можно синтезировать простые и более сложные правила сегментации фразы. Хотя существуют отдельные случаи, для которых точная сегментация затруднительна, существует возможность сегментации слитной последовательности цифр с полной ошибкой менее 1%, т. е. имеется менее 1% случаев, в которых определение границ цифр на слух показывает, что при автоматической сегментации часть сигнала данной цифры включена в сигнал другой цифры или, наоборот, к сигналу данной цифры добавлена часть сигнала следующей.

На рис. 9.27 и 9.28 показаны два примера последовательностей цифр, сегментированных системой распознавания. На этих рисунках через *а*, *б*, *в*, *г* обозначены траектории числа переходов через нуль, логарифма энергии, статистического параметра, на основе которого проводятся классификация «вокализованный—невокализованный—пауза» и принятие решения. Статистический параметр, используемый для классификации типа сегмента и представленный на рисунках кривыми *в*, означает вероятность того, что классификатор выносит правильное решение, поэтому он изменяется от нуля до единицы. Траектория *г* признака «вокализованный—невокализованный—пауза» является трехуровневой, где уровень *1* соответствует паузе, уровень *2* — невокализованному сигналу, а уровень *3* — вокализованному сигналу.

На рис. 9.27 показаны результаты сегментации последовательности цифр /721/. Первая граница расположена на начальном участке невокализованного сегмента, т. е. /s/ в слове «seven», следующая граница — в начале следующего невокализованного сегмента, соответствующего звуку /t/ в слове «two», а третья граница — в области локального минимума логарифма энергии внутри второй вокализованной зоны. Точная граница не совпадает с локальным минимумом логарифма энергии, но расположена вблизи этого минимума. Она определена с помощью ряда совместных решений алгоритма сегментации. Третья граница точно не определена, но для распознавания цифр ее точное положение внутри вокализованно-

го сегмента и не требуется. Граница последней цифры расположена в начале последней паузы. Следует отметить, что другое возможное расположение границы на рис. 9.27 соответствует точному локальному минимуму траектории логарифма энергии в звуке /v/

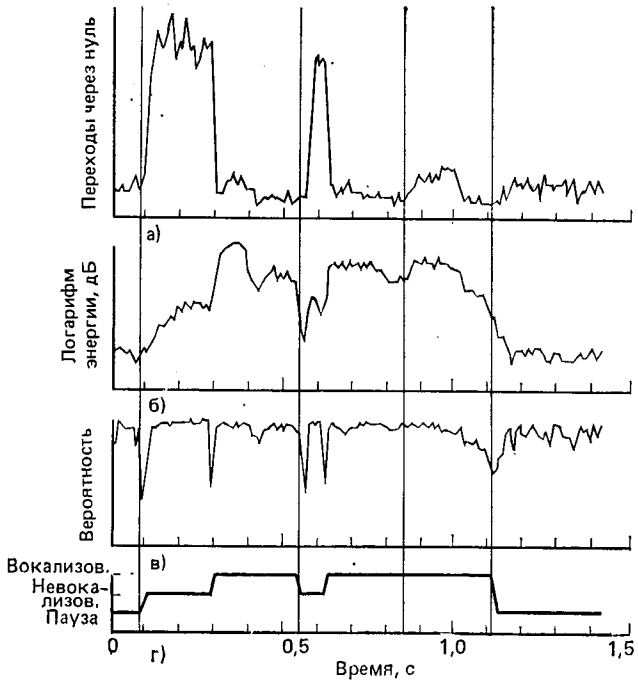


Рис. 9.27. Сегментация последовательности цифр /721/ [27]

слова «seven». Однако правило сегментации исключает этот минимум, он заменяется минимумом на следующем вокализованном сегменте.

На рис. 9.28 представлен более сложный для сегментации отрезок, соответствующий последовательности цифр /191/. Входной сигнал полностью вокализован, поэтому нет невокализованных граничных сегментов. Кроме того, отсутствуют четкие минимумы траектории логарифма энергии, т. е. эти минимумы недостаточно глубокие и имеют большую протяженность. Таким образом, расположение границ выбрано с использованием алгоритма сегментации на основе логических правил. Прослушивания показали, что расположение границ на полностью вокализованных сегментах некритично, что объясняется наличием значительной коартикуляции при произнесении таких полностью вокализованных последовательностей цифр.



Следующим шагом после сегментации является реализация алгоритма распознавания. Для каждого сегмента цифры вокализованный участок (который определяется по признаку «вокализованный—невокализованный—пауза») подвергается анализу на

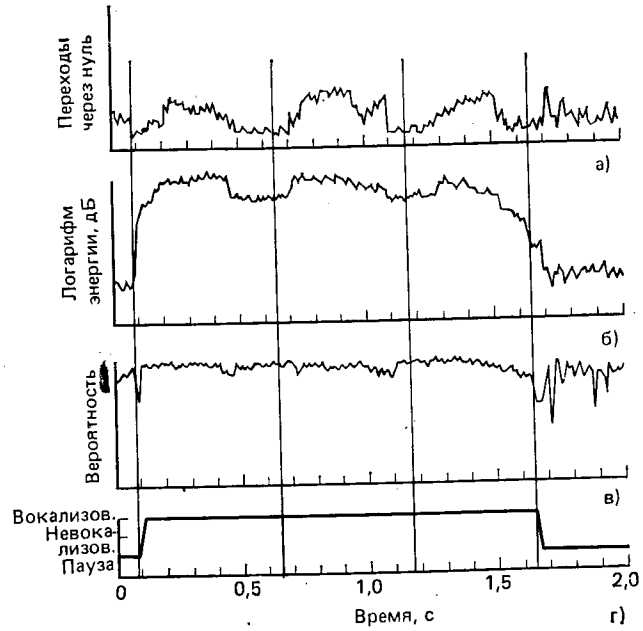


Рис. 9.28. Сегментация последовательности цифр /191/ [27]

основе линейного предсказания с использованием десятиполюсной модели. Метод распознавания основан на статистическом решающем правиле, по которому для каждого интервала обрабатываемой фразы (совокупности параметров линейного предсказания) проводится сравнение с соответствующим эталоном и выносится решение о соответствии обрабатываемой цифры той из эталонных, с которой имеется наибольшее сходство (наименьшее расстояние).

Эталонные файлы содержат статистическую информацию о коэффициентах линейного предсказания на каждом интервале анализа и для каждой цифры. В этих файлах содержится информация о среднем и дисперсии по множеству произнесений и множеству дикторов. При распознавании обрабатываемая фраза подвергается преобразованию масштаба времени выравниванием длительностей опознаваемых и эталонных цифр. Все методы преобразования временно́го масштаба, рассмотренные в § 9.1, пригодны и в данном случае. Поскольку все цифры (за исключением 7) моносиллабические, то в большинстве систем распознавания используется линейное преобразование масштаба. Для каждой эталонной фразы вычисляется среднее расстояние между ее коэффициента-

ми линейного предсказания и соответствующими коэффициентами обрабатываемой фразы, а за произнесенную принимается та цифра, для которой это расстояние оказывается минимальным. Выбор меры различимости, определяющей расстояние между совокупностями параметров линейного предсказания, является одним из наиболее важных факторов, определяющих качество работы подобных систем. Известен ряд мер различимости, используемых при обработке параметров линейного предсказания. В следующем разделе рассматриваются некоторые из них и обсуждается связь между их свойствами и статистическими свойствами параметров линейного предсказания.

Испытания систем, аналогичных представленной на рис. 9.26, показали, что, если система настроена на определенного диктора, точность распознавания достигает 98—100%, а для произвольного диктора точность составляет около 95% [27, 28].

### 9.3.3. Меры различимости в пространстве параметров линейного предсказания

Для систем распознавания диктора и речи требуется количественно и достаточно эффективно с вычислительной точки зрения сравнивать два сегмента речевого сигнала, имеющих различные коэффициенты линейного предсказания. Таким образом, необходима мера различимости $D(\mathbf{a}, \hat{\mathbf{a}})$, где $D$ — расстояние между сегментами речи с параметрами линейного предсказания $\mathbf{a}=(1, a(1), a(2), \ldots, a(p))$ и $\hat{\mathbf{a}}=(1, \hat{a}(1), \hat{a}(2), \ldots, \hat{a}(p))$. Поскольку $D$ — расстояние, следует потребовать, чтобы

$$D(\hat{\mathbf{a}}, \mathbf{a}) \geqslant 0 \text{ и } D(\hat{\mathbf{a}}, \mathbf{a}) = 0 \text{ при } \mathbf{a} = \hat{\mathbf{a}}. \qquad (9.17); (9.18)$$

Одну из таких мер различимости $D(\hat{\mathbf{a}}, \mathbf{a})$ предложил Итакура [17]. Эта мера может быть получена на основе следующих рассуждений. Предположим, что вследствие шума и неполной адекватности модели линейного предсказания речи невозможно точно оценить коэффициенты линейного предсказания на соответствующем сегменте сигнала. Оценки (измеренные значения) можно получить лишь приближенно. Предположим, что имеется сегмент сигнала с оценками параметров предсказания $\hat{\mathbf{a}}$. Задача состоит в определении вероятности того, что коэффициенты $\hat{\mathbf{a}}$ являются оценками, полученными на сегменте с истинными параметрами $\mathbf{a}$. Если такая вероятность оценена, то можно получить эффективную меру различимости сегментов.

Манн и Вальд [29] показали, что оценки $\hat{\mathbf{a}}$ распределены по многомерному закону Гаусса со средним $\mathbf{a}$ и ковариационной матрицей $\mathbf{\Lambda}$, определяемой выражением

$$\mathbf{\Lambda} = (\mathbf{R}^{-1}/N)(\hat{\mathbf{a}}\,\mathbf{R}\,\hat{\mathbf{a}}^t), \qquad (9.19)$$



где $\mathbf{R}$ — корреляционная матрица речевого сигнала размером $(p+1)\times(p+1)$; $N$ — протяженность интервала оценивания (числе отсчетов); индекс $t$ означает транспонирование. Таким образом, вероятность получения оценки $\hat{\mathbf{a}}$ при условии, что коэффициенты линейного предсказания соответствуют истинному сигналу $\mathbf{a}$ имеет вид

$$P(\hat{\mathbf{a}}/\mathbf{a}) = [(2\pi)^{p/2}\,|\mathbf{\Lambda}|^{1/2}]^{-1} \exp[-0{,}5\,(\hat{\mathbf{a}}-\mathbf{a})\,\mathbf{\Lambda}^{-1}\,(\hat{\mathbf{a}}-\mathbf{a})^t], \qquad (9.20)$$

где $|\mathbf{\Lambda}|$ — определитель матрицы $\mathbf{\Lambda}$. Соответствующая мера различимости получается, если вычислить логарифмы выражения (9.20) и пренебречь смещением за счет $|\mathbf{\Lambda}|$. Окончательное выражение для меры различимости имеет вид

$$D(\hat{\mathbf{a}}, \mathbf{a}) = (\hat{\mathbf{a}}-\mathbf{a})\left(N\frac{\mathbf{R}}{\hat{\mathbf{a}}\,\mathbf{R}\,\hat{\mathbf{a}}^t}\right)(\hat{\mathbf{a}}-\mathbf{a})^t. \qquad (9.21)$$

Чем больше вероятность того, что $\hat{\mathbf{a}}$ получено из распределения с истинными параметрами $\mathbf{a}$, тем меньше расстояние, вычисленное с использованием меры различимости (9.21). С целью преодоления вычислительных трудностей Итакура предложил весьма похожую меру различимости

$$D'(\hat{\mathbf{a}}, \mathbf{a}) = \log\left(\frac{\mathbf{a}\,\mathbf{R}\,\mathbf{a}^t}{\hat{\mathbf{a}}\,\mathbf{R}\,\hat{\mathbf{a}}^t}\right). \qquad (9.22)$$

Предложение, лежащее в основе проведенного анализа, заключается в том, что одним и тем же звукам речи соответствуют одни и те же параметры линейного предсказания на любом сегменте. Различия в оценках параметров линейного предсказания для таких сегментов целиком относят за счет статистической природы обрабатываемого сигнала. Для большого числа систем такое предположение вполне справедливо. Однако в том случае, когда коэффициенты линейного предсказания изменяются вследствие некоторых эффектов, таких, как замена диктора, коартикуляция и т. д., истинные параметры также меняются. Эти изменения лучше описывать через статистическое распределение с некоторым средним значением.

Таким образом, для полного описания некоторого сегмента речевого сигнала необходимо определить распределение $\mathbf{a}$. Целесообразно предположить, что $\mathbf{a}$ является гауссовым со средним значением $\mathbf{m}$ и ковариационной матрицей $\mathbf{S}$. На основе такого описания $\mathbf{a}$ расстояние между $\hat{\mathbf{a}}$ и $\mathbf{a}$ определяется выражением

$$\hat{D}(\hat{\mathbf{a}}, \mathbf{a}) = (\hat{\mathbf{a}}-\mathbf{m})\,\mathbf{C}^{-1}\,(\hat{\mathbf{a}}-\mathbf{m})^t, \qquad (9.23)$$

где $\mathbf{C}$ — полная ковариационная матрица вида

$$\mathbf{C} = \mathbf{S} + (R^{-1}/N)(\hat{\mathbf{a}}\,\mathbf{R}\,\hat{\mathbf{a}}^t). \qquad (9.24)$$

Для использования меры различимости (9.23) необходимо оценить величины $\mathbf{m}$ и $\mathbf{S}$ для каждого интервала анализа и каждого

эталона. Величина $m=(1, m(1), m(2), ..., m(p))$ является средним значением **a** и определяется выражением

$$m(n)=\frac{1}{y}\sum_{j=1}\hat{a}_j(n),\ n=1, 2, \ldots, p, \qquad (9.25)$$

где $\hat{a}_j(n)$, $j=1, 2, ..., j$ — оценки параметров из выборки с одним и тем же распределением **a**. Аналогично ковариационная матрица **S** с элементами $s(n, p)$ имеет вид

$$s(n, p)=\frac{1}{y}\sum_{j=1}\hat{a}_j(n)\hat{a}_j(p)-m(n)m(p). \qquad (9.26)$$

### 9.3.4. Система распознавания с большим объемом словаря

Третья из описываемых здесь систем распознавания обладает словарем, объем которого значительно превосходит объемы словарей двух первых систем. Однако платой за увеличение объема словаря является то, что система перестает быть не зависимой от диктора, т. е. система должна быть предварительно обучена применительно к каждому предполагаемому пользователю. С учетом обсуждения, проведенного во введении, разработанная Итакурой [17] система с большим словарем обладает следующими свойствами:

1. Словарь состоит из изолированных слов, количество которых составляет 100—500.
2. Система предназначена для одного диктора, но после соответствующего обучения может быть настроена на любого диктора.
3. Отсутствуют ограничения на пол и возраст диктора.
4. Отсутствуют жесткие ограничения на условия произнесения.
5. Система работает с сигналом телефонного качества.
6. Предусмотрено обучение системы в виде одно- или многократного произнесения каждого слова словаря.
7. Форматом произнесения являются слова, разделенные паузами.

На рис. 9.29 представлена структурная схема обработки сигнала в системе распознавания слов. Для повышения эффективности и снижения объема вычислений Итакура использовал частоту дискретизации 6,67 кГц. Поскольку полоса частот входного сигнала составляет 3 кГц, такая частота дискретизации вполне подходит для данного случая.

После определения моментов начала и окончания слов на основе использования методов обработки во временной области (см. гл. 4) оцениваются первые восемь коэффициентов корреляции со скоростью 67 раз/с. Для компенсации искажений спектра, вносимых телефонной линией, Итакура вычислял спектр, усредненный на большом интервале времени, что достигалось усреднением коэффициентов корреляции по всей фразе и подгонкой к усредненному по фразе спектру двухполюсной модели. Параметр-



ры двухполюсной модели использовались для построения обратного фильтра. Средний по фразе спектр затем нормировался по входу путем свертки исходных автокорреляционных коэффициентов и коэффициентов корреляции импульсной характеристики обратного фильтра. Первые шесть нормированных автокорреляцион-

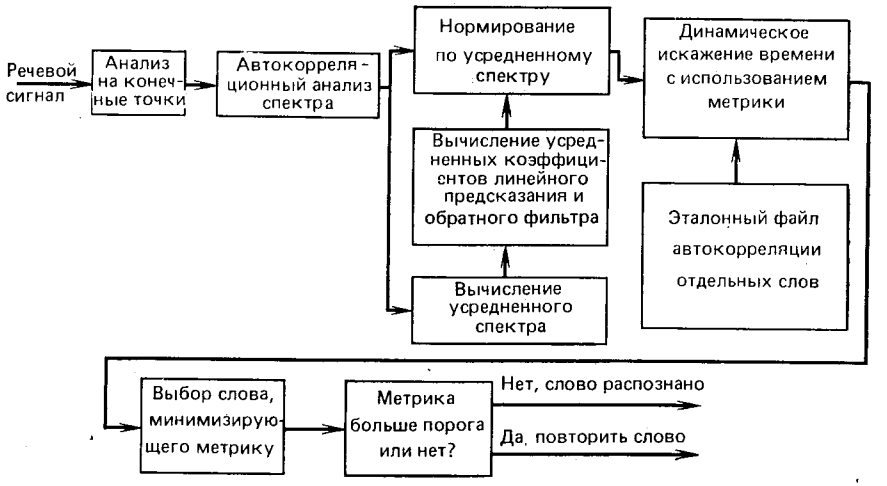


Рис. 9.29. Структурная схема системы распознавания слов, снабженная большим словарем и ориентированная на отдельного диктора

ных коэффициентов использовались затем как для создания эталонных образов, так и для распознавания.

После нормализации спектра начинается процедура распознавания. Неизвестная фраза сравнивается с каждым эталоном из имеющихся в файле. Сравнение происходит на основе меры различимости [см. 9.3.3, ф-ла (9.22)]. Эта мера использовалась также и для динамического согласования временно́го масштаба входной фразы при минимизации расстояния с каждой из эталонных фраз. На основе вычисления расстояний до каждого слова из каталога эталонов выбирается то слово, для которого полученное расстояние минимально. Если абсолютное значение расстояния превышает некоторый порог, то решение не принимается. В этом случае выбирается другое слово с минимальным расстоянием, оно принимается в качестве решения и поступает на выход системы распознавания.

Эта система исследовалась с использованием двух различных словарей. Применяя словарь объемом примерно 120 слов (названия различных городов Японии), Итакура получил частость правильного распознавания, равную 97,3%, а частость отклонения предъявленных слов 1,65%. Для словаря, соответствующего 26 буквам и цифрам от 0 до 9, полученная частость правильного распознавания равна 88,6%. Такое увеличение частости ошибок

(11,4% при частости отклонения 0%) обусловлено большим сходством между некоторыми элементами словаря, например *b* и *d*, *m* и *n* или *i* и *y*.

## 9.4. Комбинированная система речевого общения с машиной

Чтобы кратко охарактеризовать метод построения системы речевого общения человека и машины в будущем, завершим эту главу элементарным описанием системы, сочетающей верификацию и идентификацию с речевым ответом. Рассмотрим, в частности, экспериментально проверенную систему, используемую при управлении полетами самолетов и продаже авиабилетов[1].

Для описания системы приведем простой «сценарий» ее работы, позволяющий понять, как она может функционировать и быть полезной в различных сферах обслуживания. Клиент авиакомпании вызывает компьютер. Отвечает система речевого ответа.

Компьютер. *Это система информации и обслуживания авиакомпании ABC. Введите, пожалуйста, номер вашего текущего счета.* (Абонент вводит, компьютер проверяет.)

Компьютер. *Доброе утро, мистер XYZ. Пожалуйста, произнесите Вашу контрольную фразу для кредитного обслуживания.* (Клиент произносит, компьютер проверяет ее.)

Компьютер. *Спасибо, мистер XYZ. Я подтверждаю ваш текущий счет. Куда вы желаете полететь?*

Клиент. *Вашингтон.*

Компьютер. *Из какого аэропорта вы желаете лететь на Вашингтон*[2]*?*

Клиент. *Ньюарк.*

Компьютер. *Какого числа вы хотели бы вылететь из Ньюарка?* (Клиент сообщает дату или говорит, что без даты.)

Компьютер. *В какое время (такого-то числа или в такой-то день) вы бы хотели уехать?* (Клиент указывает время.)

Компьютер. *Минуточку, я сейчас дам вам номера рейсов самолетов, которые вылетают в пределах двух часов от указанного времени.* (Компьютер анализирует файл с расписанием, записанным на дисках, выбирает требуемые рейсы и зачитывает их.)

Компьютер. *Какой рейс вы желаете заказать?*

Клиент. *Рейс 123.*

Компьютер. *Вы желаете первый класс или дешевле рейсом 123?*

Клиент. *Первым классом.*

Компьютер. *Сколько мест первого класса вам нужно?*

Клиент. *Одно.*

Компьютер. *Минуточку, пожалуйста. Я проверю наличие одного места.* (Проверяет.)

Компьютер. *Я с удовольствием сообщаю* (повторяет все сначала). *Ваш билет появится на вашем телетайпе в течение часа. Счет будет прислан вам к концу месяца. Желаете ли вы заказать еще что-либо?*

Клиент. *Нет.*

Компьютер. *Благодарим за пользование услугами авиакомпании ABC.* (Отключается.)

Вариант этой системы проходит проверку в лабораториях Белла в коммутируемой сети с обусловленным заранее множеством телефонов и локальным обменом.

[1] Пример взят из статьи Дж. Л. Фланагана.

[2] Отметим два обстоятельства: 1. На каждом шаге ответное сообщение является одновременным подтверждением распознавания на последнем шаге. Если последняя фраза распознана неправильно, то клиент нажимает кнопку и система повторяет последний вопрос. 2. Словарь для распознавания на каждом шаге соответствует множеству возможных ответов на поставленный вопрос. Такая организация словаря существенно повышает устойчивость системы распознавания.



## 9.5. Заключение

В данной главе предпринята попытка иллюстрировать, каким образом методы цифровой обработки сигналов, рассмотренные в книге, могут быть использованы для построения интересных и полезных систем обработки речи. Чем больше известно о методах цифровой обработки речевых сигналов, тем более сложные системы обработки речевых сигналов можно создать и тем более широкое применение найдут полученные знания как в обычных системах передачи информации, так и в системах общения между человеком и машиной.

## Курсовые проекты

Ниже предлагаются темы курсовых проектов по цифровой обработке речевых сигналов для трех основных направлений:

I. Литературные обзоры и доклады.

II. Проекты по реализации.

III. Проекты по моделированию.

**I.** Литературные обзоры и доклады

Студент должен выбрать тему и рассмотреть следующие вопросы:

1. Суть проблемы.
2. Что является наиболее важным в этой проблеме, например, области применения и т. д.
3. В чем состоит основной подход?
4. Что уже достигнуто в данной области?
5. Необходимы ли новые подходы?
6. Какие проблемы не решены? Что требует дальнейшей проработки?
7. Что необходимо для дальнейшего прогресса, например, технология, необходимость новых фундаментальных исследований и т. д.

Предлагается несколько тем для литературного обзора:

1. Методы выделения основного тона.
2. Методы классификации речи на вокализованную и невокализованную.
3. Влияние телефонного канала на анализ речи.
4. Описание фонетических особенностей английского языка.
5. Фонетические характеристики и моделирование источников звуков для речеобразования.
6. Методы формантного анализа.
7. Синтез речи по правилам.
8. Методы адаптивного квантования.

9. Методы анализа площади поперечного сечения речевого тракта.
10. Методы идентификации дикторов.
11. Машинные системы речевого ответа.
12. Распознавание цифр с помощью ЭВМ.
13. Передача речи в гелиевой среде.
14. Системы обучения глухих речи.
15. Проблема подавления реверберации.
16. Методы подавления отражений.
17. Системы синтеза на основе коэффициентов линейного предсказания.
18. Линейное предсказание и методы идентификации систем.
19. Применение гомоморфной обработки речи.
20. Ускорение и замедление речи.
21. Нуль-полюсный анализ речи.
22. Методы анализа через синтез при обработке речи.
23. Артикуляторная модель речи.
24. Реализация методов кодирования речевой волны.
25. Системы сокращения полосы частот сигнала речи.

II. Проекты по технической реализации

Такие проекты, если это возможно, должны быть доведены до стадии технической реализации или хотя бы до стадии логических схем. Основные вопросы, решаемые в проектах такого типа, состоят в следующем:

1. В чем заключается выбранная Вами задача? Заметим, что проекты такого типа позволят Вам проявить свою изобретательность и придумать новое и более удачное решение какой-либо задачи, которая может быть уже решена.

2. Чем Вы располагаете для решения данной проблемы, например теорией и технологией?

3. В чем заключаются тонкости предлагаемого решения? Это должно быть сделано настолько подробно, насколько это возможно в пределах имеющегося времени.

4. Было ли возможным предлагаемое решение ранее? Если нет, то почему?

5. Какие требования по технической реализации необходимы для внедрения системы?

6. Желательно оценить сложность реализации (в виде количества умножителей, сумматоров, микропроцессоров, памяти и других средств хранения) и примерную стоимость устройства. Темы, предлагаемые для проектов по реализации:

1. Разработать преобразователь ИКМ в АРИКМ, ИКМ в АДМ и т. д.
2. Разработать устройство выделения основного тона.
3. Предложить систему обработки речевых сигналов, которую можно было бы реализовать на широко распространенных микропроцессорах.
4. Разработать устройство обнаружения речевого сигнала в зашумленном телефонном канале.
5. Разработать четырехполосный анализатор спектра речевого сигнала.
6. Разработать систему отображения спектрограмм речи.
7. Разработать параллельный формантный речевой анализатор.
8. Разработать устройство шифрования речи.
9. Разработать цифровое устройство выделения основного тона.
10. Разработать устройство для обнаружения вокализованной и невокализованной речи.
11. Разработать устройство различения речевого сигнала от шума.

III. Проекты по моделированию

Студентам следует браться за эти проекты, если они достаточно хорошо владеют моделированием на ЭВМ и могут получить достаточное количество машинного времени для их реализации. В рамках проектов этого типа требуется кратко описать проблему, включая математическую теорию и цели исследования, распечатки программ (с соответствующей документацией и комментариями), а также результаты контрольного счета. Для этого типа проектов предлагаются следующие темы:



1. Устройства выделения основного тона во временной области (автокорреляционные, кепстральные, на основе линейного предсказания и т. д.).
2. Устройства анализа речи на вокализованную и невокализованную.
3. Устройства определения начала и конца элементов речи.
4. Формантные анализаторы.
5. Системы анализа на основе линейного предсказания — преобразователь сигнала в спектр модели линейного предсказания.
6. $N$-канальный спектральный анализатор — фазовый и спектральный вокодеры.
7. Кодеры речевого колебания, т. е. АРИКМ (адаптивная разностная ИКМ), АДМ (адаптивная дельта-модуляция) и т. д.
8. Расчет функции площади поперечного сечения голосового тракта.
9. Исследование влияния формы и протяженности временного окна на энергию, автокорреляцию и спектрограмму речевого сигнала.
10. Речевые синтезаторы: спектральные, параллельные, прямые, лестничные.
11. Программа преобразования функции площади поперечного сечения в формантные частоты.
12. Кодопреобразователь между двумя любыми кодовыми форматами.
13. Кепстрально сглаженный спектр для сигнала речи.
14. Преобразование параметров линейного предсказания в другие параметры и исследование их спектральных свойств.
15. Сравнение спектров линейного предсказания, кепстрального и БПФ.