

А.В. Аграновский, Д.А. Леднов

**ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ АЛГОРИТМОВ
ОБРАБОТКИ И КЛАССИФИКАЦИИ
РЕЧЕВЫХ СИГНАЛОВ**

Москва
Издательство «Радио и связь»
2004

УДК 621.391
ББК 32.973
А-41

Рецензенты:

доктор физико-математических наук, профессор В.С. Пилиди,
доктор физико-математических наук, профессор Ю.В. Дацко

Аграновский А.В., Леднов Д.А.

А-41 Теоретические аспекты алгоритмов обработки и классификации речевых сигналов. Москва: Изд-во «Радио и связь», 2004. 164 с.

1133069

ISBN 5-256-01743-8

В книге рассматриваются методы обработки цифровой речи, предназначенные для формирования последовательности векторов признаков и два типа задач классификации речевого сигнала: распознавание слитной речи, идентификация диктора по его голосу. В задаче формирования векторов признаков основное внимание уделяется методам обнаружения и фильтрации речи в условиях стационарного и нестационарного шума, в том числе речеподобного, а так же методам выделения мгновенной частоты основного тона и спектральных оценок речи, которые формируют вектор акустических признаков. В задачах классификации речевых сообщений рассматриваются аспекты известной скрытой модели Маркова в сочетании с гауссовыми смесями, и способы построения различных представлений апостериорной вероятности сигнала.

Д-01 (03)-2004. Без объявл.

ISBN 5-256-01743-8

ББК 32.973

© А.В. Аграновский, Д.А. Леднов, 2004
© Издательство «Радио и связь», 2004

Содержание

ВВЕДЕНИЕ 6

Глава 1. ОБНАРУЖЕНИЕ, ФИЛЬТРАЦИЯ И ПАРАМЕТРИЗАЦИЯ РЕЧЕВОГО СИГНАЛА 13

1.1. Обнаружение речевого сигнала	13
1.1.1. Обнаружение в условиях стационарного шума	15
1.1.2. Обнаружение речи в условиях марковского шума	21
1.2. Фильтрация речевого сигнала и его восстановление ...	28
1.2.1. Фильтрация в условиях стационарного шума	29
1.2.2. Фильтрация речи в условиях нестационарного шума	31
1.2.2.1. Фильтрация импульсных шумов	32
1.2.2.2. Декомпозиция двух речеподобных сигналов..	35
1.2.2.3. Восстановление клиппированного сигнала ...	47
1.2.2.4. Восстановление потерянных блоков данных	50
1.3. Выделение основного тона	54
1.4. Параметризация и нормализация речевого сигнала.	
Оценка значимости спектральных компонент	63
1.4.1. Вычисление логарифмически масштабированных кепстральных коэффициентов	65
1.4.2. Вычисление масштабированных коэффициентов линейного предсказания	67
1.4.3. Селекция значимых компонент спектральной оценки	71

Глава 2. ИДЕНТИФИКАЦИЯ ДИКТОРОВ 77

2.1. Алгоритмы текстонезависимой системы идентификации диктора (ТСИД) для произвольных типов сегментации	81
2.1.1. Векторное квантование	81
2.1.1.1. Решающие правила	83

<i>2.1.1.2. Алгоритм обучения на основе k-средних, ЛБГ алгоритм</i>	84
<i>2.1.1.3. Метод обучения векторного квантования (OBK)</i>	
<i>2.1.2. Гауссовые смеси</i>	
<i>2.1.2.1. Решающее правило</i>	
<i>2.1.3. Методы нормализации характеристик</i>	92
2.2. Алгоритмы ТСИД с распознаванием гласных звуков	94
<i>2.2.1. Оценка точности ТСИД с распознаванием гласных звуков</i>	96
<i>2.2.2. Обучение на диктора</i>	105
<i>2.2.3. Решающее правило и тестирование ТСИД</i>	107
Глава 3. АЛГОРИТМЫ РАСПОЗНАВАНИЯ РЕЧИ	109
3.1. СММ модели фонем, трифонов и пентафонов	114
3.2. Использование нормального распределения с линейной авторегрессией для аппроксимации апостериорной вероятности	120
3.3. Использование нормальных распределений с квадратичной регрессией для аппроксимации апостериорной вероятности	124
3.4. Модель языка	127
3.5. Иерархические СММ-модели	129
ПРИЛОЖЕНИЯ	134
ЛИТЕРАТУРА	157

Обозначения и сокращения

АЦП — аналогово-цифровой преобразователь
 АЧХ — амплитудно-частотная характеристика
 ОТ — основной тон
 ДПФ — дискретное преобразование Фурье
 ОДПФ — обратное дискретное преобразование Фурье
 СММ — скрытая модель Маркова
 EM-алгоритм — Expectation Maximization algorithm
 МГУА — метод группового учета аргументов
 CASA — Computational Auditory Scene Analysis
 MFCC — Mel frequency cepstral coefficients
 RASTA — RelAtive SpecTrAl Technique

ВВЕДЕНИЕ

Математический аппарат цифровой обработки сигналов стал частью практически любого научного исследования, связанного с измерительным процессом. Как правило, под обработкой сигнала понимают решение следующих основных задач:

- создание модели сигнала;
- определение параметров модели сигнала;
- обнаружение сигнала на фоне помех;
- выделение полезного сигнала из его смеси с шумом;
- преобразование сигнала из одного представления в другое;
- определение взаимозависимостей между компонентами сигнала;
- выделение значимых компонент сигнала;
- классификация сигналов.

Специфика сигналов, которая определена физической природой их появления, и цель, которая должна быть достигнута в результате обработки сигнала, определяют выбор методов для решения той или иной задачи.

Специфика речевого сигнала связана с тем, что он порожден сложно устроенным акустическим волноводом (речевым трактом), с изменяющимися во времени геометрическими формами, а основные цели его обработки, вызванные практическими потребностями, могут быть отражены целым списком:

1. Уплотнение речевого сигнала для передачи в канале (например, телефонном) с минимальной потерей информативности [1].
2. Искажение голоса диктора при условии сохранения разборчивости его речи [2].
3. Сегментация беседы множества дикторов на монологи [3].
4. Декомпозиция беседы дикторов, т.е. расслоение сигналов по принадлежности диктору при условии их одновременного разговора [4].
5. Выделение характерных признаков голоса диктора и последующая его идентификация [5, 6].

6. Оценка эмоционального состояния диктора по его голосу [7].
7. Распознавание речи [8, 9].

8. Идентификация языка, на котором говорит диктор [10].

Последние шесть задач можно назвать *задачами классификации речевых сигналов*, так как они связаны с принятием решения о принадлежности сигнала или участка сигнала тому или иному множеству (классу).

При решении любой из перечисленных задач классификации, возникает проблема, связанная с выбором вектора признаков, которым будет характеризоваться акустические колебания. Выбор тех или иных признаков определяется как знаниями физиологии органов слуха и психоакустическими данными, так и требованием согласованности последующей модели принятия решения и вектора признаков.

Для разъяснения участия процедур формирования признаков в общей структуре систем классификации речевых сигналов рассмотрим одну из распространенных функциональных схем системы, решающей эти задачи (рис. 1).

Устройство ввода звукового сигнала — это программно-аппаратное средство измерения уровня звукового давления и его оцифровки. Как правило, такое средство состоит из микрофона, передающего канала и АЦП. (Отметим, что в дальнейшем рассматриваются только оцифрованные сигналы).

Следующий блок рисунка 1 — «Детектор речи». Он предназначен для выделения из звукового потока сегментов, содержащих речь. Необходимость использования такого блока связана с двумя аспектами:

а) звуковой поток в значительной степени состоит из пауз между словами и фразами, которые необходимо удалить, так как они не несут никакой значимой информации (для примера, в задачах идентификации диктора и распознавании речи). Удаление пауз из звукового потока сокращает время обработки, затрачиваемое последующими блоками;

б) в некоторых задачах обработки речи является информативной длительность паузы (например, в задаче оценки эмоци-

нального состояния диктора [7]), и детектор речи позволяет ее определить, т.е. детектор речи способен формировать некоторые компоненты вектора признаков.

Основным источником методов, использующихся в детекторах речи, является теория обнаружения стохастических сигналов, основные элементы которой описаны в [11,12].

Блок фильтрации связан с необходимостью удалять из речи различные помехи и искажения, источником которых является как среда передачи сигнала, так и ошибки человека при произношении того или иного звука.



Рис. 1. Структура классификатора речевых данных

Блок предварительной обработки речи предназначен для формирования последовательности векторов признаков спектрального характера и для нормализации последовательности этих векторов относительно темпа речи или длины речевого тракта диктора (необходимость применения нормализаций зависит от типа классификационной задачи). Каждый вектор признаков соответствует некоторому сегменту речи, длительность которого выбрана исследователем из соображений целесообразности. В общем случае признаки сегмента состоят из следующего набора:

- мощность речевого сигнала (абсолютная мощность и ее производные);
- основной тона (мгновенная частота, ее производные, показатели асимметрии импульса основного тона);
- спектральные показатели (абсолютные значения спектральных компонент и их производные).

Такой состав вектора признаков выбран потому, что исследования акустических свойств речевого тракта показали [13, 14], что результирующий спектр акустического сигнала, создаваемого речевым трактом, и получаемый нашими органами восприятия, можно представить в виде:

$$f(\omega) = A(\omega)T(\omega)F(\omega)R(\omega) + \eta(\omega),$$

где ω — частота; $A(\omega)$ — спектр огибающей, формируемой деятельностью легких; $T(\omega)$ — спектр основного тона, который определяется работой голосовых складок; $F(\omega)$ — передаточная характеристика речевого тракта; $R(\omega)$ — функция, учитывающая частотную зависимость, обусловленную активной составляющей сопротивления излучению; $\eta(\omega)$ — спектр шума, который может быть образован как шумами окружающей среды, так и шумами самого человеческого тела.

Классификация методов спектральной обработки, которые принято называть методами параметризации сигнала, приведена на рисунке 2. Здесь необходимо отметить, что такие классификации очень быстро устаревают. Так, в аналогичном перечне методов, приведенном в работе [15], отсутствовала ветвь, связанная с вейв-

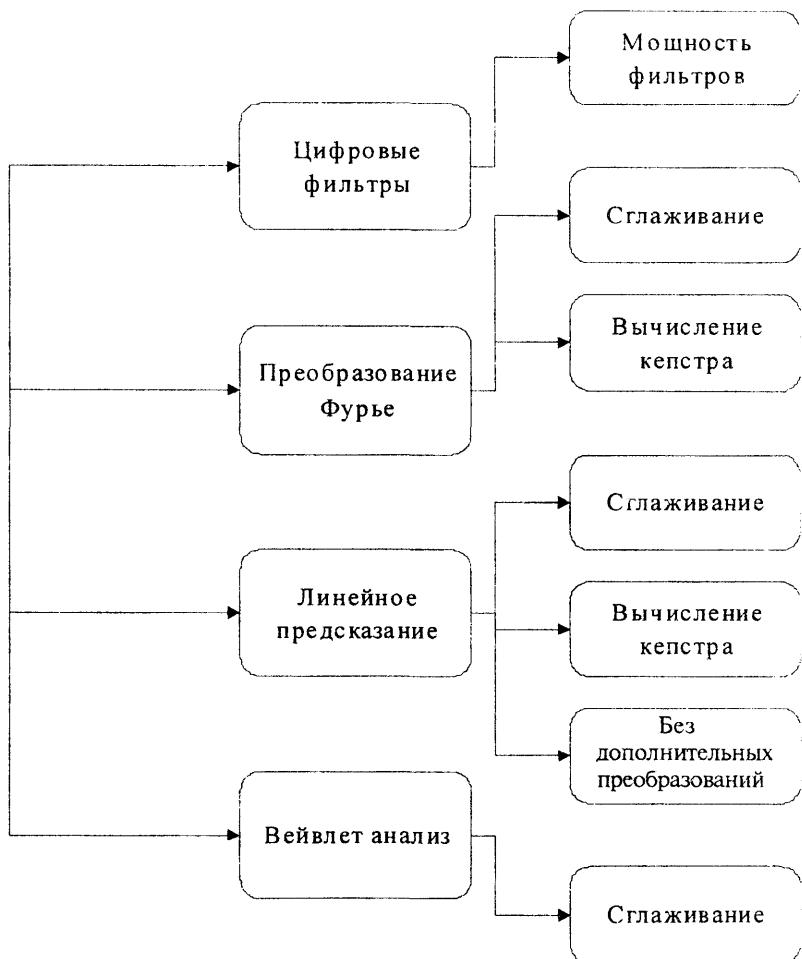


Рис. 2. Классификация методов предварительной обработки сигнала

лет анализом, который сейчас активно используется в системах обработки речи.

Алгоритмы обучения (см. рис. 1) предназначены для установления соответствия между входной последовательностью векторов признаков и последовательностью наименований классов. Под наименованием класса может подразумеваться имя дикто-

ра, эмоциональное состояние диктора, слово, фонема и т.д. (в зависимости от цели обработки речи).

Задача, которая решается алгоритмами классификации (рис. 1), является обратной по отношению к задаче обучения, т.е. здесь требуется по входной последовательности векторов признаков найти соответствующую последовательность наименований классов.

Скрытая модель Маркова (СММ) является наиболее теоретически развитой, используемой для построения алгоритмов обучения и классификации речевых сообщений. Она описана во многих отечественных и зарубежных публикациях, например [15–17]. В приложении 1 мы приведем ее краткое описание.

В рамках представленной функциональной схемы системы классификации речевых сигналов (рис. 1) мы видим несколько проблем, которые рассмотрим в этой работе совместно с классическими подходами.

Проблема первая. Детектирование и фильтрация речи в условиях нестационарных шумов, которые существенно влияют на качество классификации речи. Здесь будут рассмотрены шумы марковского, импульсного и речеподобного характера.

Проблема вторая. Алгоритмы обучения и распознавания должны быть согласованы с вектором признаков. Согласованность в современных системах классификации заключается в том, что в любой момент времени вектор признаков принадлежит пространству R' , т.е. размерность вектора не изменяется со временем. Это в свою очередь означает, что всякий сегмент сигнала описывается одинаковым количеством информации вне зависимости от содержания этого сегмента. Такой недостаток векторов признаков может быть преодолен введением размерности в зависимости от содержания. Эта процедура будет рассмотрена на примере использования формантных характеристик речи и введения способа сравнения двух формантных наборов.

Проблема третья. При идентификации дикторов не используется способ по фонемной идентификации. Конечно, такой способ приводит к необходимости использовать совместно модель распознавания фонем и модель идентификации дикторов

и требует больших затрат времени, но это наша плата за точность, которую мы можем достигнуть.

Проблема четвертая. В работах по распознаванию слитной речи наблюдается тенденция роста контекстной зависимости для более точного моделирования звучания той или иной фонемы в окружении других фонем. Такой контекст можно назвать фонетическим. На этом пути были развиты СММ для фонем, трифонов и пентафонов. Чем выше фонетическая контекстная зависимость, тем больше затраты памяти вычислительной системы и объем выборки речи, необходимый для обучения системы: если система, основанная на трифонной модели, имеет около 75 000 состояний, то система, основанная на пентафонной модели, уже 130 700 000 состояний. Однако можно предложить путь, основанный на акустическом контексте, где предполагается зависимость текущего вектора признаков от множества окружающих его соседей. Объем памяти, требующийся для этой модели значительно меньше, и введение изменений в используемую длительность контекста значительно проще. Эти модели будут рассмотрены при аппроксимации апостериорных вероятностей для СММ.

Первая и третья главы, где рассмотрены вопросы, связанные с обнаружением сигнала в шуме, фильтрацией и распознаванием слитной речи, были написаны Д.А. Ледновым. Глава вторая (о методах идентификации диктора) — А.В. Аграновским.

Авторы выражают благодарность сотрудникам Лаборатории обработки речи НИИ «Спецвузавтоматика» г. Ростова-на-Дону за внимательное прочтение работы и сделанные ценные замечания.

Глава 1. ОБНАРУЖЕНИЕ, ФИЛЬТРАЦИЯ И ПАРАМЕТРИЗАЦИЯ РЕЧЕВОГО СИГНАЛА

1.1. Обнаружение речевого сигнала

Детектор речи предназначен для выделения из входного звукового потока, состоящего из смеси полезного сигнала и шума, непрерывной последовательности сегментов, содержащих законченную фразу или слово. Задача детектирования речи сходна с задачей, решаемой в рамках классической теории обнаружения стохастических сигналов, которая описана во многих публикациях (например, [11, 18]). Здесь кратко изложим основные положения этой теории.

Пусть $\eta_t = \eta(t\Delta t)$ — значение аддитивного шума в дискретный момент времени t , полученное в результате оцифровки АЦП аналогового шума через равные интервалы времени Δt , ξ_t — значение полезного сигнала (речи) в дискретный момент времени t , тогда y_t — значение наблюдаемого сигнала может быть записано в виде

$$y_t = \theta\xi_t + \eta_t, \quad (1.1)$$

где θ — случайная величина, определенная на множестве $\{0,1\}$; $\theta=1$ в случае если наблюдаемый сигнал содержит полезную составляющую, и $\theta=0$ в обратном случае.

По значению наблюдаемого сигнала нам необходимо принять решения о наличии в нем полезной составляющей. Таких решений может быть два: полезного сигнала нет: $\delta_0=0$, и полезный сигнал есть: $\delta_1=1$. Пространство реализаций Ω наблюдаемого сигнала должно быть разбито на два подпространства, каждое из которых соответствует определенному решению. Пусть подпространство Ω_0 соответствует решению δ_0 , а пространство Ω_1 решению δ_1 . Однако принимаемые решения могут быть ошибочными, и для того чтобы характеризовать эти ошибки вводится показатель риска

$$R = (1 - p_1)K(0, \delta_1) \int_{\Omega_1} p(y|0)dy + p_1 K(1, \delta_0) \int_{\Omega_0} p(y|1)dy, \quad (1.2)$$

где p_1 — априорная вероятность наличия полезного сигнала в наблюдаемом процессе; $p(y|\theta)$ — условная плотность вероятности наблюдаемого сигнала в зависимости от случайной величины θ ; в качестве весов этих вероятностей выступает положительно определенная функция потерь $K(\theta, \delta)$, выбираемая эмпирически. Показатель риска равен сумме взвешенной вероятности принять решение δ_1 , при условии, что полезного сигнала нет (первый член суммы в (1.2)) и взвешенной вероятности принять решение δ_0 , при условии, что полезный сигнал присутствует (второй член суммы в (1.2)). В нашем случае функцию потерь удобно записать в матричной форме

$$K(\theta, \delta) = K_{\theta\delta} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Учитывая, что $\Omega_1 = \Omega / \Omega_0$ и $\int_{\Omega} p(y|\theta) dy = 1$, показатель риска (1.2) можно записать в виде

$$R = (1 - p_1)K(0, \delta_1) + \int_{\Omega_0} (p_1 K(1, \delta_0) p(y|1) - (1 - p_1) K(0, \delta_1) p(y|0)) dy. \quad (1.3)$$

Отсюда следует, что риск минимален, если подынтегральное выражение в (1.3) равно нулю. Таким образом, решение принимает вид

$$\delta = \begin{cases} 1, & \text{если } \lambda > \mu, \\ 0, & \text{если } \lambda \leq \mu, \end{cases} \quad (1.4)$$

где введены обозначения $\lambda = \frac{p(y|1)}{p(y|0)}$; $\mu = \frac{(1 - p_1)K(0, \delta_1)}{p_1 K(1, \delta_0)}$.

Величину λ обычно называют отношением правдоподобия, а способ получения решения (1.4) — байесовским.

Далее, на основе решения Байеса (1.4) рассмотрим методы построения детектора речи для различных разновидностей шумовой обстановки.

1.1.1. Обнаружение в условиях стационарного шума

Как известно [19], случайный процесс y_t , определенный на числовом множестве V , называется стационарным, если $M\{y_t^2\} < \infty$, где $M\{\cdot\}$ — операция вычисления математического ожидания, математическое ожидание $m_t = M\{y_t\}$ не зависит от времени, а корреляционная функция $R_{tt} = M\{(y_t - m_t)(y_t - m_t)\}$ зависит лишь от разности аргументов $R_{tt} = R_{t-t}$.

В реальных условиях измерения случайных процессов воспользоваться приведенным определением невозможно, так как количество реализаций случайного процесса и их длительность ограничены. Этот факт заставляет использовать оценки величин [20], приведенных в определении.

Рассмотрим соотношение между оценками случайного процесса во временной и частотной областях. Известно, что несмещенной оценкой математического ожидания случайного процесса y_t является величина

$$m_t = \frac{1}{T} \sum_{i=t}^{t+T-1} y_i.$$

Найдем дискретное преобразование Фурье (ДПФ) случайного процесса на интервале T (допустим, что это возможно):

$$m_t = \frac{1}{T} \sum_{i=t}^{t+T-1} \sum_{k=0}^{T-1} a_{kt} W^{ki} = \frac{1}{T} \sum_{k=0}^{T-1} a_{kt} W^{tk} \sum_{i=t}^{t+T-1} W^{ik} = \frac{1}{T} \sum_{k=0}^{T-1} a_{kt} W^{tk} \frac{1 - W^{kT}}{1 - W^k}, \quad (1.5)$$

где $W = \exp\{j \frac{2\pi}{T}\}$; a_{kt} — k -ый коэффициент ДПФ, полученный в дискретный момент времени t .

Очевидно, что дробь в последнем выражении (1.5) равна нулю за исключением случая $k=0$, откуда заключаем, что оценка математического ожидания всегда равна нулевому коэффициенту ДПФ случайного процесса, т.е. $m_t = a_{0t}$ и если случайный процесс стационарен, то $a_{0t} = \text{const}, \forall t$.

Далее докажем утверждение, что если процесс стационарен, то оценка математического ожидания любой компоненты спект-

ра — постоянная величина. Для этого проведем вычисления, соответствующие ОДПФ компонент случайного процесса:

$$\begin{aligned} h_{kd} &= \frac{1}{T} \sum_{t=d}^{d+T-1} a_{kt} a_{kt}^* = \frac{1}{T} \sum_{t=d}^{d+T-1} \sum_{g=0}^{T-1} y_{g+t} W^{-k(g+t)} \sum_{l=0}^{T-1} y_{l+t} W^{k(l+t)} = \\ &= \frac{1}{T} \sum_{t=d}^{d+T-1} \sum_{g=0}^{T-1} \sum_{l=0}^{T-1} y_{g+t} y_{l+t} W^{k(g-l)} = \sum_{\tau=0}^T R_\tau W^{k\tau}. \end{aligned} \quad (1.6)$$

Вычисления показывают, что независимость корреляционной функции от времени обеспечивает справедливость утверждения.

Выражение (1.6) позволяет проводить исследование стационарного случайного процесса не во временной области, а в частотной, что практически более удобно, так как появляется возможность игнорировать фазы гармоник, которые не влияют на восприятие звуков речи человеком.

Речь не является стационарным процессом с точки зрения определения, приведенного в начале этого параграфа, так как и математическое ожидание и корреляционная функция являются зависимыми от времени. Однако в некоторых случаях (при определенных взаимоотношениях между спектральными компонентами шума и речи) для решения задачи обнаружения возможно рассматривать ее как стационарный процесс.

Рассмотрим поведение нестационарного полезного сигнала в евклидовом пространстве спектральных компонент. Динамика этих компонент отображается траекторией, которая может быть целиком помещена в некоторую замкнутую область Q . Траектория стационарного шума так же может быть помещена в замкнутую область q . Если области Q и q пересекаются, то речь можно описать некоторым распределением вероятности, присущим стационарному процессу.

Предположим, что плотность распределения вероятности спектральных компонент как шума, так и полезного сигнала имеют гауссову форму

$$p(h_0, h_1, \dots, h_{T-1}; \mathbf{m}, D) = \frac{1}{\sqrt{(2\pi)^T \det D}} \exp\left\{-\frac{1}{2}(\mathbf{h} - \mathbf{m})D^{-1}(\mathbf{h} - \mathbf{m})\right\}, \quad (1.7)$$

где h_i — значение i -ой спектральной компоненты; \mathbf{m} — вектор математических ожиданий спектральных компонент;

$$m_i = \frac{1}{T} \sum_{t=0}^T h_{it}; \quad (1.8)$$

D — корреляционная матрица, элементы которой определены как

$$D_{ij} = M\{(h_{it} - m_i)(h_{jt} - m_j)\}. \quad (1.9)$$

Классифицировать тот или иной входной сигнал, в смысле (1.1), который в момент времени t обладает вектором спектральных компонент \mathbf{h}_t , возможно с помощью байесовского решения (1.4), аргументы которого могут быть записаны в виде

$$\lambda = \frac{p(h_0, h_1, \dots, h_{T-1} | 1; \mathbf{m}^{(\xi+\eta)}, D^{(\xi+\eta)})}{p(h_0, h_1, \dots, h_{T-1} | 0; \mathbf{m}^{(\eta)}, D^{(\eta)})}, \quad (1.10)$$

$$\mu = 1$$

где $(\mathbf{m}^{(\eta)}, D^{(\eta)})$, $(\mathbf{m}^{(\eta+\xi)}, D^{(\eta+\xi)})$ — параметры распределения плотности вероятности, соответственно шума и смеси шума с полезным сигналом, и принято, по отношению к (1.4), что априорная вероятность наличия полезного сигнала $p_1 = 1/2$ и значения функции потерь равны $K(0,1) = K(1,0) = 1$.

После того, как выбрана вероятностная модель сигнала (1.7) и стратегия принятия решения (1.10), необходимо оценить параметры распределений (1.8) и (1.9). Практически это означает, что требуется создать две достаточно представительные выборки случайных процессов, первая из которых должна содержать только шум, а вторая — только смесь полезного сигнала и шума. Конечно, такую работу можно провести «вручную», т.е. опытный оператор проведет сортировку записей речевых сообщений и выделит два необходимых класса выборок, но более предпочтителен вариант автоматического создания таких выборок. Для автоматизации процесса мы должны построить более простой, но достаточно точный метод классификации сигнала.

Поскольку в качестве модели случайного процесса выбрана

его аддитивная форма (1.1), то это позволяет предполагать, что среднее значение квадрата амплитуды смеси полезного сигнала и шума будет превышать среднее значение квадрата амплитуды шума.

Сигнал, содержащий только шум, как правило, получить не представляет сложности. Например, в системах, где обеспечен ввод речи с помощью микрофона, внутренний шум системы и внешний могут быть записаны в режиме молчания диктора, а в системах, где ввод обеспечен телефонным каналом, есть достаточно время от момента снятия трубки до начала беседы (около 0,5 с).

Предположим, что среднее значение квадрата амплитуды шума на интервале T

$$A_t = \frac{1}{T} \sum_{d=t}^{d+T-1} \eta_d^2 \quad (1.11)$$

распределено по нормальному закону

$$p(A_t; \theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(A_t - \theta)^2}{2\sigma^2}\right\}, \quad (1.12)$$

где θ — оценка математического ожидания средних квадратов амплитуды

$$\theta = \frac{1}{G} \sum_{i=0}^{G-1} A_i; \quad (1.13)$$

σ — дисперсия

$$\sigma^2 = \frac{1}{G} \sum_{i=0}^{G-1} (A_i - \theta)^2, \quad (1.14)$$

G — количество интервалов длительностью T , выбранных для оценки параметров распределения (1.12).

Классифицируем сегменты, содержащие полезный речевой сигнал с помощью условия $p(A_t; \theta, \sigma) < H$, где H — порог, которое при логарифмировании приводит к следующему неравенству

$$A_t - \theta > \sigma \sqrt{\ln(2\pi^2 H^2)}. \quad (1.15)$$

Таким образом, для построения детектора речи на фоне стационарного шума необходимо выполнить следующие операции:

1. Создать запись, содержащую только шум.
2. На основе этой записи рассчитать параметры распределения (1.12), т.е. вычислить дисперсию и математическое ожидание по формулам (1.13) и (1.14).
3. На основе этой же записи оценить параметры распределения (1.7) для шума, т.е. вычислить (1.8) и (1.9). На рисунке 1.1. показано математическое ожидание (сплошная линия) и сумма математического ожидания и двукратной дисперсии (пунктирная линия) распределения (1.7) с диагональной корреляционной матрицей для шума компьютерного вентилятора.

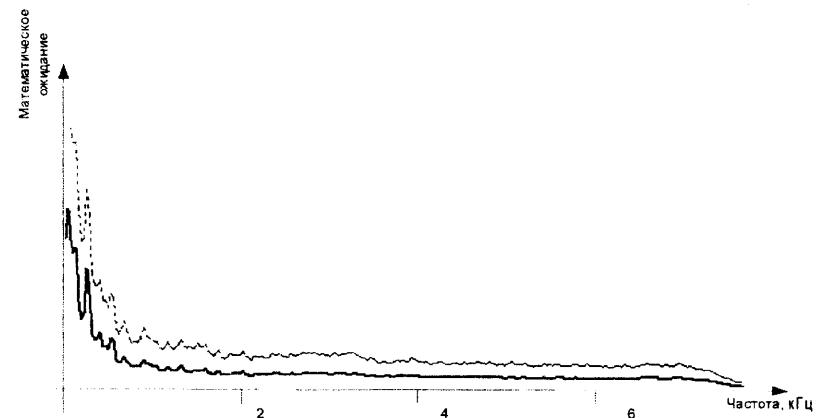


Рис. 1.1. Распределение вероятностей спектральных компонент шума компьютерного вентилятора

4. Создать запись, содержащую речь.
5. Классифицировать сегменты этой записи с помощью условия (1.15).
6. На основе сегментов классифицированных как смесь полезного сигнала и шума оценить параметры распределения (1.7). На рисунке 1.2. показано математическое ожидание (сплошная линия) и сумма математического ожидания и двукратной дис-

персии (пунктирная линия) распределения (1.7) с диагональной корреляционной матрицей для речи.

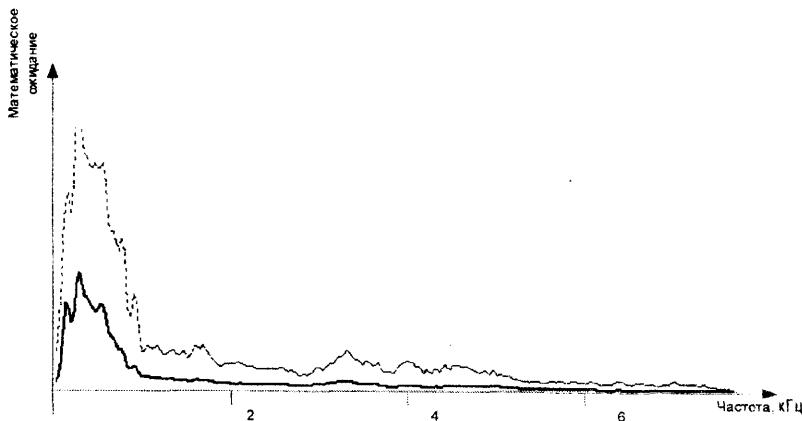


Рис. 1.2. Распределение вероятностей спектральных компонент речи

7. Классифицировать новый входной сигнал с помощью решения Байеса с аргументами (1.10).

Здесь была описана одна из простейших моделей детекции речи в условиях стационарного шума, отличающаяся последовательным использованием амплитудных и частотных принципов обнаружения сигнала. Очевидно, что этот класс моделей является перспективным, а его развитие может быть направлено по двум путям:

а) поиск распределений вероятностей, которые могут более точно аппроксимировать данные об амплитудах компонент, как спектра шума, так и спектра речи. Для примера, распределение Гаусса в формуле (1.7) может быть замещено или смесью гауссоид, параметры которой можно найти с помощью известного EM-алгоритма [21], описанного в приложении 2, или нормальным процессом авторегрессии [17] с алгоритмом обучения, приведенным в работе [22] (см. раздел п.1.1.2.);

б) выделение значимых компонент спектра шума и их функций, наиболее достоверно позволяющих отличить шум от по-

лезного сигнала. Эту операцию можно выполнить с помощью МГУА [23], который также будет описан ниже.

1.1.2. Обнаружение речи в условиях марковского шума

В отличие от постановки задачи, описанной в предыдущем параграфе, где нам были известны распределения шума и смеси полезного сигнала и шума, здесь рассмотрим случай, когда известны отдельно статистические свойства речи и статистические свойства шума, а наблюдается их смесь. В качестве статистических моделей шума и речи используем СММ [17].

В дополнении к обозначениям и определениям, принятым для СММ в приложении 1, введем обозначения, которые нам потребуются в рамках этого параграфа:

- \mathbf{h}_t — вектор наблюдения, составленный из компонент спектра Фурье входного сигнала, $\mathbf{h}_t \in R^n$, $\forall t$, причем $\mathbf{h}_t = \theta \chi_t + \mathbf{z}_t$, здесь χ_t , \mathbf{z}_t — вектора компонент спектра Фурье полезного сигнала и шума, соответственно, θ — определена в соответствии с (1.1);
- $Z = \{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{G-1}\}$ — предполагаемая последовательность значений шума;
- $U = \{u_0, u_1, \dots, u_{G-1}\}$ — некоторая произвольная последовательность состояний полезного сигнала;
- $V = \{v_0, v_1, \dots, v_{G-1}\}$ — некоторая произвольная последовательность состояний шума;

В дальнейшем для упрощения записи вероятностей будем использовать только индексы текущих состояний последовательности.

Рассмотрим два способа обнаружения полезного сигнала. Первый из них основан на использовании байесовского решения (1.4), второй — на построении наиболее вероятной траектории на обобщенном множестве состояний смеси полезного сигнала и шума и чистого шума.

Первый способ. Будем предполагать, что априорная вероятность наличия полезного сигнала $p_1=1/2$ и значения функции потерь равны $K(0,1)=K(1,0)=1$ соответственно $\mu=1$, а для отношения правдоподобия справедливо

$$\lambda = \frac{\max_{U,V,Z} p(H, Z, U, V | 1)}{\max_{V^*} p(H, V^* | 0)} = \frac{\max_{U,V,Z} p(U, V | 1) p(U, V | H, Z; 1)}{\max_V p(V^* | 0) p(V^* | H; 0)} \quad (1.16)$$

где $p(U, V | 1)$ — вероятность последовательности смеси состояний шума и полезного сигнала. Очевидно, что в силу их независимости справедливо

$$p(U, V | 1) = p(V | 1) p(U | 1) = p(u_0) p(v_0) \prod_{t=1}^G p(u_t | u_{t-1}) p(v_t | v_{t-1}), \quad (1.17)$$

где $p(U, V | H; 1)$ — условная вероятность последовательности смеси состояний шума и полезного сигнала при реализации последовательности наблюдений H . Опять же в силу независимости состояний сигнала и шума и при условии, что известны значения шума в каждый момент времени, для нее справедливо

$$p(U, V | H, Z; 1) = \prod_{t=0}^G p(u_t | \mathbf{h}_t - \mathbf{z}_t) p(v_t | \mathbf{z}_t), \quad (1.18)$$

где $p(V^* | 0)$ — вероятность последовательности состояний шума V^* (звездочкой отмечено то, что состояния чистого шума не совпадают с его состояниями в смеси); $p(V^* | H; 0)$ — вероятность последовательности состояний шума V^* при наблюдении последовательности H .

Раскроем выражение (1.16), используя (1.17) и (1.18):

$$\lambda = \frac{\max_{U,V,Z} \pi(u_0, v_0 | \mathbf{h}_0, \mathbf{z}_0) \prod_{t=1}^G \psi(u_t, v_t | \mathbf{h}_t, \mathbf{z}_t)}{\max_{V^*} p(v_0^*) p(v_0^* | \mathbf{h}_0) \prod_{t=1}^G p(v_t^* | \mathbf{h}_t) p(v_t^* | v_{t-1}^*)}, \quad (1.19)$$

где для сокращения записи введены обозначения

$$\pi(u_0, v_0 | \mathbf{h}_0, \mathbf{z}_0) = p(u_0) p(u_0 | \mathbf{h}_0 - \mathbf{z}_0) p(v_0) p(v_0 | \mathbf{z}_0),$$

$$\psi(u_t, v_t | \mathbf{h}_t, \mathbf{z}_t) = p(u_t | u_{t-1}) p(u_t | \mathbf{h}_t - \mathbf{z}_t) p(v_t | v_{t-1}) p(v_t | \mathbf{z}_t).$$

Отношение правдоподобия (1.19) имеет очень высокую вычислительную сложность за счет многомерной максимизации, и его практически невозможно применить в таком виде. Однако, можно использовать два пути выхода из сложившейся ситуации: 1) провести операцию суммирования состояний шума и полезного сигнала и за счет этой операции снять необходимость максимизации знаменателя (1.19) по двум параметрам из трех; 2) предположить, что диагональные элементы матриц переходных вероятностей много больше недиагональных и воспользоваться методом кумулятивных сумм [17].

Рассмотрим вариант упрощения вычислений в (1.19) за счет суммирования состояний шума и полезного сигнала.

Пусть $S_s = \{s\}$ — множество состояний полезного сигнала, содержащее их ровно N_s штук, а $S_d = \{d\}$ — множество состояний шума, содержащее N_d состояний, тогда множество смешанных состояний $S = S_s \oplus S_d = \{f\}$ будет содержать $N = N_d N_s$ новых состояний. Очевидно, что для матрицы переходов между элементами множества S и условных вероятностей справедливы соответствующие соотношения

$$p(s_i + d_k | s_j + d_l) = p(f_{i(N_d-1)+k} | f_{j(N_d-1)+l}) = p(s_i | s_j) p(d_k | d_l),$$

$$p(s_i + d_j | \mathbf{h}) = p(f_{i(N_d-1)+j} | \mathbf{h}) = \int_{R''} p(s_i | \mathbf{h} - \mathbf{z}) p(d_j | \mathbf{z}) d\mathbf{z}, \quad (1.20)$$

(здесь мы вводим перенумерацию смешанных состояний, чтобы в дальнейшем пользоваться для их обозначения одним индексом).

Если обозначить в качестве $W = \{w_0, w_1, \dots, w_{G-1}\}$ последовательность смешанных состояний, то используя соотношения (1.20), отношение правдоподобия (1.19) можно записать в виде

$$\lambda = \frac{\max_w p(w_0) p(w_0 | \mathbf{h}_0) \prod_{t=1}^G p(w_t | \mathbf{h}_t) p(w_t | w_{t-1})}{\max_{V^*} p(v_0^*) p(v_0^* | \mathbf{h}_0) \prod_{t=1}^G p(v_t^* | \mathbf{h}_t) p(v_t^* | v_{t-1}^*)}, \quad (1.21)$$

Поиск максимума числителя или знаменателя заключается в том, чтобы среди всевозможных последовательностей состоя-

ний найти такую последовательность, которая бы с максимальной вероятностью описывала наблюдаемый процесс. Будем называть такую последовательность НВ-траекторией. Вычислить ее можно, используя метод динамического программирования [17], описанный в приложении 1.

Несмотря на значительное упрощение отношения правдоподобия (1.19), за счет построения множества смешанных состояний, и приведение его к виду (1.21), само вычисление значения (1.21) по прежнему остается трудоемким из-за двукратного использования метода динамического программирования (приложение 1). При условиях жестких требований ко времени обнаружения такой метод неприменим. Более простым является способ, использующий механизм разладки, впервые предложенный в работе [24]. Суть этого метода состоит в том, чтобы в каждый момент времени вычислять логарифм коэффициента правдоподобия в виде

$$\lambda_t = \ln \frac{\max_{i=1..N} p(f_i | \mathbf{h}_t)}{\max_{i=1..N_d} p(d_i | \mathbf{h}_t)}. \quad (1.22)$$

Затем рассчитывать кумулятивную сумму логарифмов коэффициентов правдоподобия по правилу

$$g_t = \begin{cases} g_{t-1} + \lambda_t & \text{при } g_{t-1} + \lambda_t > 0, \\ 0 & \text{при } g_{t-1} + \lambda_t \leq 0 \end{cases}, \quad (1.23)$$

и следить за ее знаком. Если знак (1.23) положительный, то наблюдается полезный сигнал, в противном случае — шум.

Как показано в [17], такой подход справедлив только в случае, если диагональные элементы матриц вероятностей переходов между состояниями много больше недиагональных.

Второй способ. Пусть имеется множество состояний случайного процесса S , которое состоит из состояний, соответствующих смеси речи и шума $f_{i(N_d-1)+j} = s_i + d_j$ и состояний чистого шума $f_{N_s N_d + j} = d_j$. Обозначим текущее состояние из этого мно-

жества как o_t , а последовательность состояний как O . Тогда вероятность, что данная последовательность состояний O определена данной последовательностью наблюдений будет равна

$$P(O|H) = p(o_0)p(o_0|\mathbf{h}_0)\prod_{t=1}^G p(o_t|\mathbf{h}_t)p(o_t|o_{t-1}). \quad (1.24)$$

Используя метод динамического программирования (приложение 1) найдем последовательность состояний O^* , которая максимизирует функционал (1.24). Для обнаружения сигнала, необходимо выделить из последовательности состояний O^* такие цепочки состояний, номера которых меньше, чем $N_s N_d - 1$.

Необходимо отметить, что матрица переходов определена отдельно для состояний с номерами меньшими $N_s N_d - 1$ и отдельно для состояний большими этого числа, т.е. вероятности переходов из состояний шума в смешанные состояния и обратно равны нулю. В этой ситуации можно только предположить, что такие переходы равновероятны. Это предположение приводит к тому, что матрица переходов пересчитывается в соответствии с нормальными стохастическими условиями (приложение 1):

$$p^*(f_i | f_j) = \frac{p(f_i | f_j) + N_d^{-1}}{2} \quad \text{при } j < N_d N_s,$$

$$p^*(f_i | f_j) = \frac{p(f_i | f_j) + (N_d N_s - 1)^{-1}}{2} \quad \text{при } j \geq N_d N_s.$$

Итак, к настоящему моменту нами были рассмотрены три метода классификации случайного процесса на два класса (шум и речь): два из них основаны на байесовском решении и один — на максимизации апостериорной вероятности. Однако для того чтобы их использовать, необходимо проделать еще несколько операций:

- 1) найти множество состояний случайного процесса;
- 2) выбрать вид функций условных плотностей распределения вероятностей $p(f_j | \mathbf{h}_t)$;
- 3) найти параметры этих функций;

4) вычислить матрицы переходных вероятностей между состояниями.

Найдем множество состояний для случайных процессов, соответствующих полезному сигналу и шуму. Будем считать, что мы обеспечены достаточным количеством выборок, для того чтобы определить статистические свойства шума. Что касается речи, то можно поступить следующим способом: провести исследование статистических свойств речи в условиях слабого стационарного шума, который может быть выделен детектором, описанном выше (раздел 1.1.1), и как в случае шума, — собрать достаточное количество выборок полезного сигнала.

С помощью ДПФ преобразуем полученные выборки в последовательности векторов акустических параметров $H = \{\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{N-1}\}$ и проведем кластеризацию векторов акустических параметров сигнала с помощью известного метода минимакс, описанного в [25]. Здесь изложим суть этого метода.

Пусть необходимо сортировать данные $H = \{\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{N-1}\}$ на M кластеров k_i , $i=1, \dots, M$ (причем M неизвестно), и задана мера удаленности двух векторов $d(\mathbf{h}_i, \mathbf{h}_j)$. В качестве меры удаленности можно выбрать евклидово расстояние, расстояние Махalanобиса [26] или др.

Проиллюстрируем этот метод, используя шесть векторов в выборке ($N=5$). На первом этапе разместим $N+1$ векторов в таблице, произвольно припишем вектору \mathbf{h}_0 кластер k_0 (рис. 1.3, *a*). Затем найдем вектор наиболее удаленный от кластера k_0 , например \mathbf{h}_4 ; припишем вектору \mathbf{h}_4 кластер k_1 (рис. 1.3, *б*). Теперь найдем кластер, ближайший к каждому из векторов и запомним эти минимальные расстояния. Найдем наибольшее из этих минимальных расстояний и отнесем соответствующий вектор к категории k_2 . Предположим, что этим вектором является вектор \mathbf{h}_5 (рис. 1.3, *в*). Теперь для остальных векторов найдем среди образованных кластеров ближайший к каждому из векторов и запомним расстояния. Найдем наибольшее из этих наименьших расстояний. На основании этого примера приходим к заключению: четвертое измеренное макси-

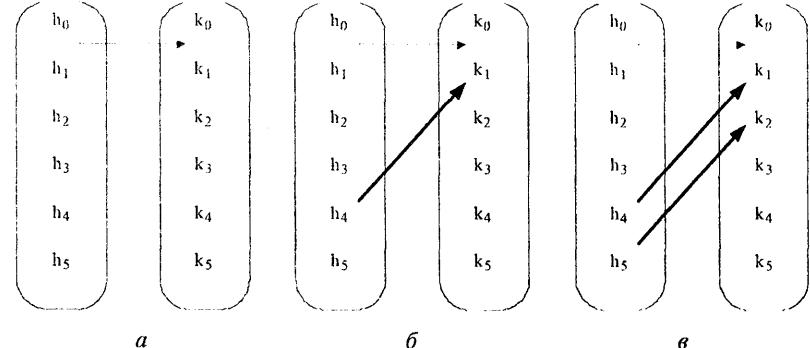


Рис. 1.3. Построение кластеров методом минимакс

мальное из минимальных расстояний существенно меньше прежнего максимального расстояния. Это показывает, что имеется три кластера. По минимальному расстоянию всех векторов, относительно образованных кластеров, возможно отнести каждый вектор к определенному кластеру.

В качестве функции условной плотности вероятности акустических параметров состояний $p(f_j | \mathbf{h}_i)$ можно выбрать функцию вида (1.7), и на основе множеств наблюдаемых векторов, отнесенных к тому или иному кластеру по формулам (1.8) и (1.9), найдем корреляционную матрицу и математическое ожидание этих распределений.

Найдем элементы матрицы переходных вероятностей $p(f_i | f_j)$. В работе [17] показано, что на основе метода обратной связи можно рассчитать эту матрицу как отношение частоты встречи состояния i после состояния j , к общей частоте встречи состояния i . Считается, что в момент времени t встречено состояние i , если выполняется условие $i = \operatorname{argmax}_j p(f_i | \mathbf{h})$.

Таким образом, для построения детектора речи на фоне нестационарного шума, имеющего Марковский характер, в соответствии с первым способом необходимо выполнить следующие операции:

1. Создать записи сигналов, содержащих только шум.

2. Создать записи, которые содержат речь в стационарном шуме с известными статистическими свойствами.

3. На базе этих записей методом минимакс определить множества состояний шума и полезного сигнала.

4. Для каждого состояния определить плотность распределения вероятностей вида (1.7) с корреляционной матрицей (1.8) и математическим ожиданием (1.9).

5. Найти значения матрицы переходов между состояниями.

6. Найти отношение между минимальным диагональным элементом матрицы вероятностей переходов и ее максимальным недиагональным элементом. Если это отношение много больше единицы, то наличие речи в поступающем на вход детектора сигнале можно определять на основе механизма разладки (1.22)–(1.23).

7. Если отношение между минимальным диагональным элементом матрицы вероятностей переходов и ее максимальным недиагональным элементом имеют один и тот же порядок, то необходимо использовать метод динамического программирования (приложение 1.2), т.е. найти оптимальную траекторию неизвестного входного сигнала на каждом интервале анализа и по формуле (1.21) определять значение коэффициента правдоподобия, который является показателем наличия речи в поступающем на вход детектора сигнале.

Как и в случае детектирования речи в стационарном шуме, представленные модели могут быть развиты в направлении поиска распределений вероятностей, которые могут более точно описывать распределения амплитуд компонент, как спектра шума, так и спектра речи в состояниях СММ.

1.2. Фильтрация речевого сигнала и его восстановление

В общем случае фильтрация сигнала состоит в том, чтобы выделить полезную составляющую сигнала и удалить посторонние шумы и искажения. В зависимости от характера шума возникает несколько задач:

а) необходимо выделить полезный сигнал из высокочастотного или полосового шума. В этом случае фильтрация заключается в выборе типа фильтра и расчете его параметров [11];

б) требуется выделить речевой сигнал из речеподобного шума [4]. Например, два или более дикторов могут говорить одновременно, а требуется получить разборчивую речь только одного из них. Это одна из самых сложных задач фильтрации, общих методов решения которой пока нет;

с) требуется восстановить сигнал, потерпевший нелинейные искажения. Эта задача возникла с появлением цифровых телефонных линий, которые уплотняют речь при передаче и исказывают исходный сигнал.

Рассмотрим основные аспекты решения этих задач.

1.2.1. Фильтрация в условиях стационарного шума

Рассмотрим систему обработки сигналов, в которой на первом шаге вычисляется ДПФ входного сигнала x_t

$$h_{in} = \sum_{t=n\Delta T}^{n\Delta T+T} W^{it} x_t, \quad (1.25)$$

где i — номер компоненты преобразования; ΔT — шаг, с которым находится преобразование; n — номер шага; T — величина окна ДПФ.

Затем каждая компонента проходит фильтрацию с помощью нерекурсивного фильтра вида

$$g_{in} = \sum_{k=0}^M a_{ik} h_{in-k}. \quad (1.26)$$

Определим, как соотносится нерекурсивная фильтрация компонент ДПФ с нерекурсивной фильтрацией сигнала во времени. Для этого синтезируем выходной сигнал y_m через компоненты g_{in} с помощью ОДПФ и затем подставим (1.25) и (1.26) в это выражение, т.е.:

$$y_m = \sum_{i=1}^{T/2} g_{in} W^{-it} = \sum_{i=0}^{T/2-1} \sum_{k=0}^M a_{ik} W^{-it} \sum_{p=(n-k)\Delta T}^{(n-k)\Delta T+T} W^{ip} x_p = \sum_{k=0}^M \sum_{p=(n-k)\Delta T}^{(n-k)\Delta T+T} S_{ikp} x_p, \quad (1.27)$$

где введено обозначение

$$S_{ikp} = \sum_{i=0}^{T/2-1} W^{-it} a_{ik} W^{ip}. \quad (1.28)$$

Из формулы (1.28) можно сделать вывод, что если для каждой компоненты ДПФ реализованы различные нерекурсивные фильтры, т.е. коэффициенты a_{ik} зависят от номера компоненты, то это равносильно тому, что во времени реализуется нерекурсивный фильтр с зависящей от времени, периодической амплитудно-частотной характеристикой (АЧХ). Величина периода АЧХ равна длительности окна ДПФ. Если же для каждой компоненты ДПФ фильтры одинаковы, то справедливо

$$S_{ikp} = a_k \sum_{i=0}^{T/2-1} W^{-it} W^{ip} = a_k \delta(p-t), \quad (1.29)$$

где $\delta(t)$ — дельта-функция Дирака. При подстановке (1.29) в (1.27) получим нерекурсивный фильтр во времени.

Таким образом, можно сделать следующие выводы: 1) нерекурсивная фильтрация компонент ДПФ входного сигнала одинаковыми фильтрами равносильна нерекурсивной фильтрации этого сигнала во времени; 2) нерекурсивная фильтрация компонент ДПФ входного сигнала произвольными фильтрами равносильна нерекурсивной фильтрации этого сигнала с помощью фильтра с зависящей от времени периодической АЧХ с периодом, равным длительности окна ДПФ.

Простейший способ синтеза нерекурсивных фильтров состоит в представлении, что нам известна достаточно представительная выборка сигнала, который необходимо отфильтровать $x_i = \xi_i + \eta_i$, состоящий из смеси полезной составляющей ξ_i и шума η_i , и дана выборка полезной составляющей сигнала. Тогда коэффициенты фильтров для каждой компоненты ДПФ можно найти с помощью метода наименьших квадратов, т.е. минимизируя функционалы

$$F_i = \sum_{n=M\Delta T}^N \left(\sum_{k=0}^M a_{ik} (h_{n-k} + \gamma_{n-k}) - h_n \right)^2, \quad \forall i = 1..T/2, \quad (1.30)$$

где γ_{in} — i -ая компонента ДПФ шума.

Поскольку дальнейшие вычисления аналогичны для всех компонент ДПФ, то игнорируем соответствующий им индекс. Дифференцируя (1.30) по коэффициентам a_m получим линейную систему алгебраических уравнений

$$\sum_{k=0}^M a_k (\theta_{km} + \phi_{km} + \phi_{mk} + \vartheta_{km}) = \theta_{0m} + \phi_m, \quad (1.31)$$

где введены обозначения

$$\theta_{km} = \sum_{n=M\Delta T}^N h_{n-k} h_{n-m}, \quad \phi_{km} = \sum_{n=M\Delta T}^N \gamma_{n-k} h_{n-m}, \quad \vartheta_{km} = \sum_{n=M\Delta T}^N \gamma_{n-k} \gamma_{n-m}.$$

Предполагая, что шум стационарный и не коррелирует с полезным сигналом, систему уравнений (1.31) можно преобразовать к виду

$$\sum_{k=0}^M a_k (\theta_{km} + \vartheta_{km}) = \theta_{0m}, \quad \text{где } m = 0..M. \quad (1.32)$$

Решение системы позволяет полностью определить фильтр.

1.2.2. Фильтрация речи в условиях нестационарного шума

В общем случае идеи фильтрации речи в условиях нестационарного шума могут развиваться в двух направлениях, которые зависят от степени наших знаний об этом шуме.

Первое направление основано на том, что информации о физическом процессе генерации нестационарного шума достаточно для построения его модели, позволяющей по некоторым параметрам отличать компоненты шума от компонент полезного сигнала.

Второе направление подходит для случая, когда такую физическую модель шума построить нельзя. Развитие этого направления основано на том, что фильтр проходит стадию обучения, где с помощью реализаций незашумленной речи создаются ее состояния. Процесс фильтрации сходен с процессом распознавания, где определяется состояние речи, с максималь-

ной вероятностью сходное с участком зашумленной речи, и которое впоследствии заменяет собой этот участок. Поскольку вариативность речи от диктора к диктору высока, то такая фильтрация предполагает либо шумоочистку сигнала заранее известного диктора, на которого и проводилось обучение фильтра, либо создание достаточно большого банка голосов дикторов и соответствующих им состояний в надежде, что удастся перебрать все типы голосов.

Здесь будут рассмотрены модели, развивающиеся в рамках первого направления для некоторых частных случаев нестационарных шумов.

1.2.2.1. Фильтрация импульсных шумов

Характеристики работы систем обработки речи во многом зависят от типа канала, по которому передается речь. Универсальный подход, основанный на оценке качества канала передачи речи с помощью соотношения сигнал/шум, не всегда оправдан. В современных типах телефонных каналов существуют классы искажений, которые, существенно не уменьшая отношения сигнал/шум, приводят к падению уровня разборчивости речи в канале. Одним из шумов такого типа является импульсный шум, физические причины которого связаны с атмосферными явлениями и кратковременными переключениями на самой автоматической телефонной станции. Пример импульсного шума приведен на рисунке 1.4.

Один из методов подавления импульсных шумов основан на свойстве нерегулярности амплитудно-частотных характеристик шума такого характера и, в противовес, на свойстве относительной регулярности амплитудно-частотных характеристик речи.

Процесс подавления шума происходит в два этапа. Первый этап связан с анализом речевого сигнала на предмет наличия в нем шума и удалением шума в случае его присутствия, а второй этап связан с синтезом сигнала без шума.

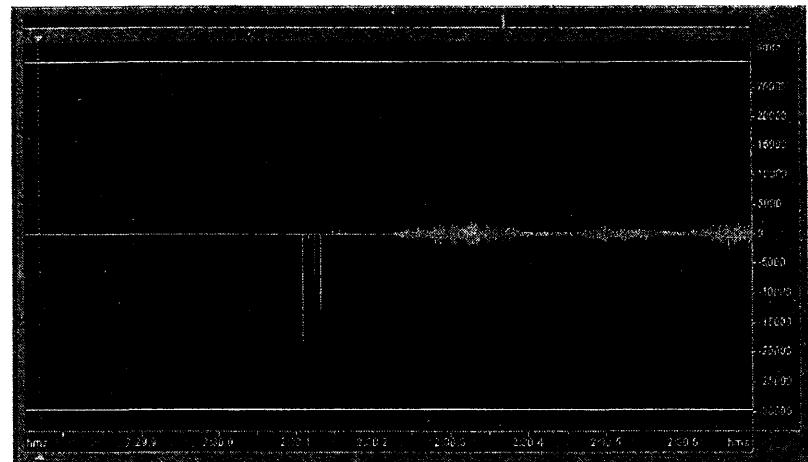


Рис. 1.4. Импульсный шум телефонного канала

Первый этап фильтрации основан на аппроксимации спектра Фурье речевого сигнала и выделении формантных наборов $L_t = \{(A, \omega)_k\}$, с помощью метода, описанного в приложении 3.

В соответствии с данным методом, проведем сравнение двух соседних наборов и выделим формантные траектории.

Если известна средняя длительность импульса, имеющего шумовое происхождение, можно установить длительность траектории параметров импульсов шума. Пусть эта длительность равна T . Далее мы можем удалить из всех траекторий, получаемых при обработке речи, траектории пар параметров, длительностью меньше, чем T .

Перейдем к рассмотрению второго этапа. Опишем один из возможных методов использования полученной информации для синтеза сигнала без шума.

Пусть для некоторого момента времени t было получено множество параметров $L_t = \{(A, \omega)_k\}_t$. После проведения описанных выше преобразований из L_t было получено множество $\tilde{L}_t = \{(A, \omega)_k\}_t$, из которого, возможно, удалены некоторые пары

$(A, \omega)_k$. Используя пары из множества L_t , возможно для любого ω найти ω_1 и ω_2 такие, что:

- 1) $\omega_1 < \omega$;
- 2) существует A_1 такое, что пара $(A_1, \omega_1) \in L_t$;
- 3) не существует пары $(A, \omega_k) \in L_t$, $\omega_k < \omega$, такой, что $\omega - \omega_k < \omega - \omega_1$;
- 4) $\omega_2 > \omega$;
- 5) существует A_2 такое, что пара $(A_2, \omega_2) \in L_t$;
- 6) не существует пары $(A_1, \omega_1) \in L_t$, $\omega_1 > \omega$ такой, что $\omega_1 - \omega < \omega_2 - \omega$.

Или, другими словами, ω_1 — ближайшая слева, а ω_2 — ближайшая справа точки к точке ω , принадлежащие множеству L_t . Построим теперь следующую функцию:

$$\Phi_t(\omega) = \begin{cases} 0 & ,(A_1, \omega_1) \notin \tilde{L}_t, (A_2, \omega_2) \notin \tilde{L}_t \\ 1 & ,(A_1, \omega_1) \in \tilde{L}_t, (A_2, \omega_2) \in \tilde{L}_t \\ h_1(\omega, \omega_1, \omega_2), & (A_1, \omega_1) \in \tilde{L}_t, (A_2, \omega_2) \notin \tilde{L}_t \\ h_2(\omega, \omega_1, \omega_2), & (A_1, \omega_1) \notin \tilde{L}_t, (A_2, \omega_2) \in \tilde{L}_t \end{cases}$$

Функции $h_1(\omega, \omega_1, \omega_2)$ и $h_2(\omega, \omega_1, \omega_2)$ выбираются из условий непрерывности функции $\Phi_t(\omega)$:

$$h_1(\omega_1, \omega_1, \omega_2) = 1, \quad h_1(\omega_2, \omega_1, \omega_2) = 0, \quad h_2(\omega_1, \omega_1, \omega_2) = 0, \quad h_2(\omega_2, \omega_1, \omega_2) = 1.$$

Например, при использовании в качестве $h_1(\omega, \omega_1, \omega_2)$ и $h_2(\omega, \omega_1, \omega_2)$ линейных функций, получится функция $\Phi_t(\omega)$, вид которой приведен на рисунке 1.5.

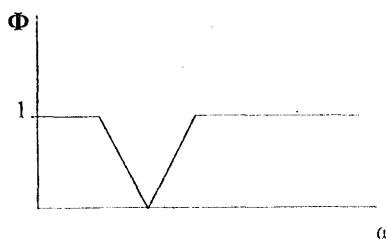


Рис. 1.5. Пример мгновенной передаточной характеристики фильтра для подавления импульсных шумов

Используя функцию $\Phi_t(\omega)$, как передаточную характеристику фильтра, вычислим из спектра входного сигнала g_j спектр полученного сигнала \tilde{g}_j :

$$|\tilde{g}_j| = |g_j| |\Phi_t(j)|, \quad \text{Arg}(\tilde{g}_j) = \text{Arg}(g_j).$$

Проводя обратное преобразование Фурье, из \tilde{g}_j получаем сигнала без импульсного шума.

1.2.2.2. Декомпозиция двух речеподобных сигналов

Зачастую в задачах обработки речи возникают события, когда два и более диктора говорят одновременно в одном передающем тракте или параллельно с беседой дикторов звучит музыкальное произведение. Здесь термин — передающий тракт можно понимать в широком смысле, т.е. это может быть акустическая среда, телефонный канал, микрофон или любое другое, пригодное для передачи речи устройство. В связи с этим возникает проблема декомпозиции (расслоения) одновременных сигналов, порожденных двумя или более источниками звука и полученными в одном передающем тракте.

Рассматриваемый класс задач возник недавно и назван *Computational Auditory Scene Analysis (CASA)* [27]. В рамках этого класса задач возникли два подхода к их решению. Один из них — *bottom-up* — предполагает проведение декомпозиции смешанных сигналов на основе только акустических данных; а другой — *top-down* — сепарацию сигналов на основе предварительного обучения и признаков информационного характера [28,29].

Основное внимание работы будет сосредоточено на трех аспектах проблемы декомпозиции:

1. Построить модель речевого тракта и выяснить, как отражается динамика речевого тракта на спектральных компонентах выходного сигнала.
2. Сформировать спектр, возникающий в результате суперпозиции колебаний от пары источников.

3. Рассмотреть алгоритм, который позволяет сепарировать два смешанных сигнала [30].

В монографии [14] были исследованы вопросы распределения давления в речевом тракте при статических геометрических формах тракта и различных граничных условиях (абсолютно твердых и податливых стенках тракта), и получены характерные спектры для большого класса вокализованных и невокализованных звуков. Автором [14] было показано, что для речевого тракта справедливо приближение «очень узких труб» и применение уравнения Вебстера:

$$\frac{\partial}{\partial t} \left(S(x,t) \frac{\partial}{\partial t} p(x,t) \right) = \frac{\partial}{\partial x} \left(S(x,t) \frac{\partial}{\partial x} p(x,t) \right), \quad (1.33)$$

где $S(s, t)$ — сечение речевого тракта, которое изменяется с координатой и течением времени; $p(x, t)$ — давление.

В отличие от [14], здесь нас будут интересовать вопросы изменения спектральных параметров звуков при изменении формы тракта, в соответствии с предположением, что именно эта динамика специфичным образом отражается в спектральной картине и позволяет сепарировать спектр, порожденный одним источником от спектра другого.

Рассмотрим модель речевого тракта, составленную из двух рупоров с абсолютно упругими стенками (рис. 1.6). У рупора $[x_1, x_2]$ в точке x_1 находится мембрана — источник колебаний вида

$$u = u_0 \exp\{-i\omega t\} / t.$$

Сечение рупоров увеличивается со временем по законам

$$S_1(x,t) = \alpha(x+x_0)^2(t+t_0)^2 \text{ и } S_2(x,t) = \beta x^2(t+t_0)^2,$$

где α, β — скорости изменения сечений первого и второго рупоров, соответственно, причем $\beta > \alpha$, t_0 — начальный момент времени.

Координата точки сочленения рупоров x_2 вычисляется из условия равенства сечений в этой точке

$$x_2 = \frac{x_0 \alpha (1 + \sqrt{\frac{\beta}{\alpha}})}{\alpha - \beta}.$$

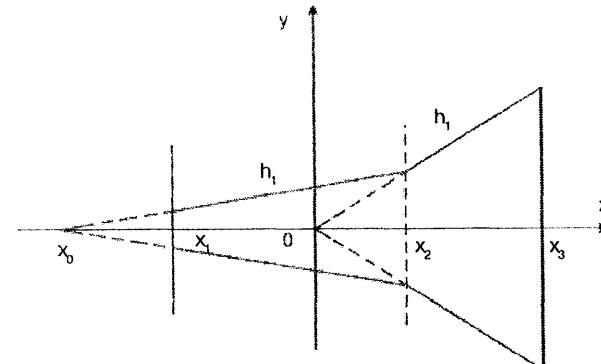


Рис. 1.6. Модельная форма речевого тракта

Уравнения Вебстера (1.33) для первого и второго рупора можно записать в виде:

$$\left(\frac{\partial^2}{\partial t^2} + \frac{2}{t+t_0} \frac{\partial}{\partial t} \right) p_1(x,t) = \left(\frac{\partial^2}{\partial x^2} + \frac{2}{x+x_0} \frac{\partial}{\partial x} \right) p_1(x,t), \quad (1.34*)$$

$$\left(\frac{\partial^2}{\partial t^2} + \frac{2}{t+t_0} \frac{\partial}{\partial t} \right) p_2(x,t) = \left(\frac{\partial^2}{\partial x^2} + \frac{2}{x+x_1} \frac{\partial}{\partial x} \right) p_2(x,t). \quad (1.34**)$$

Границные условия для этой системы уравнений

$$\begin{aligned} \left. \frac{\partial u}{\partial t} \right|_{x=x_0} &= -\frac{1}{\rho} \left. \frac{\partial p_1}{\partial x} \right|_{x=x_0}, \quad p_1(x_1,t) = p_2(x_1,t), \\ \left. \frac{\partial p_1(x,t)}{\partial x} \right|_{x=x_1} &= \left. \frac{\partial p_2(x,t)}{\partial x} \right|_{x=x_1}, \quad p_2(x_2,t) = 0. \end{aligned} \quad (1.35)$$

Решения уравнений (1.34) можно получить методом разделения переменных, который приводит их к паре обыкновенных дифференциальных уравнений с известными решениями:

$$p_1(x,t) = \frac{e^{i\omega t}}{(x+x_0)(t+t_0)} (a_1 e^{ikx} + b_1 e^{-ikx}), \quad (1.36*)$$

$$p_1(x,t) = \frac{e^{i\omega t}}{x(t+t_0)} (a_2 e^{ikx} + b_2 e^{-ikx}). \quad (1.36**)$$

Известно, что АЧХ волновода можно рассчитать по формуле [31]:

$$A(\omega) = \frac{x_1^4 |u|^2}{x_3^2 |u_3|^2}, \quad (1.37)$$

где u_3 — скорость колебаний элементов воздуха с координатой x_3 . Получить аналитическое выражение для АЧХ (1.37) трудно, поскольку они для свободных коэффициентов решений (1.36), которые определяются граничными условиями (1.35), очень громоздки. Воспользуемся численным расчетом АЧХ, результаты которого приведены на рисунке 1.7.

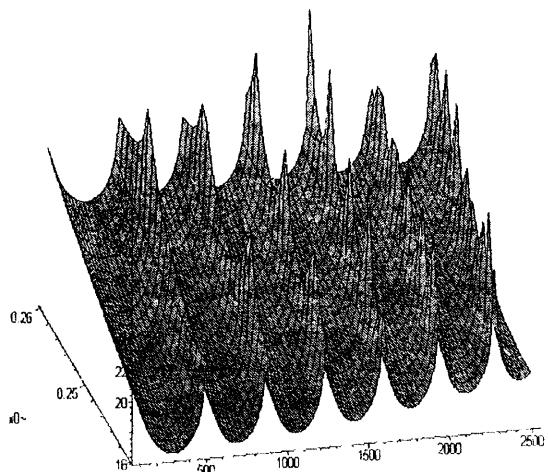


Рис. 1.7. Динамика АЧХ волновода при изменении его геометрической формы. Здесь приведена функция $\ln(A(\omega))$

При увеличении длины речевого тракта максимумы АЧХ сменяются в область высоких частот. Модуляция максимумов, которая видна на рисунке 1.7 — это лишь эффект построения графика на дискретной сетке с равным шагом. Построение этой

же функции с большей точностью приводит к исчезновению модуляции.

Качественно полученный результат понятен, так как с изменением геометрической формы волновода меняется его АЧХ. Однако изменение формы АЧХ — это не единственный источник изменений, которые наблюдаются в спектре речевого сигнала. Другим источником является основной тон.

Известно, что речевой тракт возбуждается с помощью основного тона $U(t)$, который генерируется голосовыми складками. Если считать основной тон периодической функцией с медленно меняющимся периодом $T(t)$, то его функцию можно представить в виде

$$U(t) = \sum_{j=0}^{T-1} a_j \exp(i\omega_j(t)t), \quad (1.38)$$

где $\omega_j(t) = \frac{2\pi}{T(t)}$, $j = \omega_0 + \varepsilon(t)$, $\varepsilon(t)$ — функция, модулирующая значения гармоник основного тона, которая зависит от формы речевого тракта.

Разлагая (1.38) в ряд по ε и удерживая члены первого порядка малости, получим форму амплитудной модуляции гармоник спектра при динамике речевого тракта:

$$b_j(t) = \varepsilon^2(t) A_j a_j.$$

Таким образом, нам качественно удалось выяснить, что амплитудная модуляция спектральных компонент определяется динамикой речевого тракта, а их частотная модуляция определена изменениями в работе голосовых складок.

Рассмотрим задачу декомпозиции смеси двух периодических сигналов с постоянным периодом. Пусть периодические сигналы в смеси заданы в общем виде:

$$f_t = \sum_{j=1}^N a_j \exp\left\{i\frac{2\pi}{G} jt\right\}, \quad (1.39)$$

где G — период функции, выраженный в количестве дискретных отсчетов; N — количество компонент разложения этой функ-

ции; t — номер отсчета; $|a_j| \neq 0, \forall j = 1..N, N \gg 1$. Рассмотрим процедуру измерения комплексных компонент спектра этой функции с произвольного номера отсчета τ с помощью ДПФ ограниченного прямоугольным окном длительностью T , которая не кратна величине G .

$$F_{q,\tau} = \sum_{t=0}^{T-1} f_{t+\tau} \exp\left\{-i\frac{2\pi}{T}tq\right\} = \frac{1}{T} \sum_{j=1}^N a_j \exp\left\{i\frac{2\pi}{T}\alpha_j t\tau\right\} \frac{1-W_j}{1-R_{jq}}, \quad (1.40)$$

где введены обозначения $\alpha = T/G$; $W_j = \exp\{i2\pi a_j\}$;

$$R_{jq} = \exp\left\{i\frac{2\pi}{T}(\alpha_j - q)\right\}, \text{ причем } a \gg 1.$$

Очевидно, что при выполнении условия приближенного равенства амплитуд спектральных компонент a_j , наибольший вклад в q -ую компоненту спектра доставляет член суммы, в последнем сомножителе которого $j = \left[\frac{q}{\alpha}\right]$, где $[.]$ — операция округления числа. Тогда для декомпозиции смеси сигналов с амплитудами спектральных компонент $\{a_j\}_N$, $\{b_j\}_M$ и относительными длительностями периода функции и длительности окна ДПФ α и $\beta = T/Q$, (Q — длительность периода другого сигнала смеси) необходимо построить множество систем уравнений

$$\begin{cases} a_{\left[\frac{q}{\alpha}\right]} A_{\left[\frac{q}{\alpha}\right]q,\tau} + b_{\left[\frac{q}{\beta}\right]} B_{\left[\frac{q}{\beta}\right]q,\tau} = TF_{q,\tau}, \\ a_{\left[\frac{q+1}{\alpha}\right]} A_{\left[\frac{q+1}{\alpha}\right]k+1,\tau} + b_{\left[\frac{q+1}{\beta}\right]} B_{\left[\frac{q+1}{\beta}\right]k+1,\tau} = TF_{q+1,\tau}, \end{cases} \quad (1.41)$$

где введены обозначения

$$A_{\left[\frac{q}{\alpha}\right]q,\tau} = \exp\left\{i\frac{2\pi}{T}\left[\frac{q}{\alpha}\right]\alpha\tau\right\} \frac{1 - \exp\left\{i2\pi\alpha\left[\frac{q}{\alpha}\right]\right\}}{1 - \exp\left\{i\frac{2\pi}{T}\left(\alpha\left[\frac{q}{\alpha}\right] - q\right)\right\}},$$

$$B_{\left[\frac{q}{\beta}\right]q,\tau} = \exp\left\{i\frac{2\pi}{T}\left[\frac{q}{\beta}\right]\beta\tau\right\} \frac{1 - \exp\left\{i2\pi\beta\left[\frac{q}{\beta}\right]\right\}}{1 - \exp\left\{i\frac{2\pi}{T}\left(\beta\left[\frac{q}{\beta}\right] - q\right)\right\}}. \quad (1.42)$$

Построить множество систем вида (1.41) можно, если выполняется условие, что для каждого номера m и n спектральных компонент из множеств $\{a_j\}_N$, $\{b_j\}_M$, найдется такое q , что

$$\left[\frac{q}{\alpha}\right] = m, \quad \left[\frac{q+1}{\alpha}\right] = m \quad \text{и} \quad \left[\frac{q}{\beta}\right] = n, \quad \left[\frac{q+1}{\beta}\right] = n. \quad \text{Эти условия равносильны условиям } \alpha, \beta > 2.$$

Для комплексно сопряженных спектральных компонент смеси можно построить аналогичное (1.41) множество систем линейных уравнений.

Для расчета коэффициентов систем линейных уравнений (1.41) требуется определить значение неизвестных α и β . Целочисленную оценку их можно сделать, если найти последовательность положений максимумов спектра $\bar{u} = \operatorname{argmax}_q |F_q|^2$. Далее, если предположить, что целочисленные значения равны v_1, v_2 и из последовательности $\bar{u} = \{u_1, u_2, \dots, u_B\}$, где B — количество максимумов в спектре, получить две новые последовательности

$$D = \left\{ \operatorname{sign}\left(\left[\frac{u_1}{v_1}\right] - \left[\frac{u_0}{v_1}\right]\right), \operatorname{sign}\left(\left[\frac{u_2}{v_1}\right] - \left[\frac{u_1}{v_1}\right]\right), \dots, \operatorname{sign}\left(\left[\frac{u_B}{v_1}\right] - \left[\frac{u_{B-1}}{v_1}\right]\right) \right\}$$

и

$$H = \left\{ \operatorname{sign}\left(\left[\frac{u_1}{v_2}\right] - \left[\frac{u_0}{v_2}\right]\right), \operatorname{sign}\left(\left[\frac{u_2}{v_2}\right] - \left[\frac{u_1}{v_2}\right]\right), \dots, \operatorname{sign}\left(\left[\frac{u_B}{v_2}\right] - \left[\frac{u_{B-1}}{v_2}\right]\right) \right\},$$

где $\operatorname{sign}(x) = \begin{cases} 1, & \text{если } x = 1, \\ 0, & \text{иначе} \end{cases}$.

Построенные последовательности позволяют классифицировать принадлежность того или иного спектрального максимума функции определенного периода. Таким образом, если спектр порожден смесью двух функций, то наилучшая оценка целочисленных периодов этих функций определяется выражением

$$\left(\sum_{i=1}^B D_i(v_1) \wedge H_i(v_2) - B \right)^2 = \min.$$

После того как проведено оценивание округленных значений α и β , можно провести уточнение их значений. Пусть округленное значение α равно v_1 , тогда соответственно этому в линейчатом спектре смеси мы должны находить максимумы на частотах sv_1 , где $s = 1..N$. Но благодаря мантиссе числа α происходит смещение линий спектра. Если максимум, который должен был находиться в позиции sv_1 , найден в позиции sv_1+1 , или в позиции sv_1-1 , тогда поправку γ легко вычислить из

$$\text{условия } s - \frac{sv}{v+\gamma} = \frac{sv \pm 1}{v+\gamma} - s, \text{ откуда следует } \gamma = \pm \frac{1}{2s}.$$

Таким образом, $\alpha = v + \gamma$. Аналогичные вычисления проводятся и для β . Подстановка значений α и β в (1.42) позволяет вычислить значение коэффициентов множества линейных уравнений (1.41) и провести декомпозицию смеси двух периодических сигналов вида (1.39). Очевидно, что приведенные рассуждения можно обобщить на случай смеси большего количества сигналов.

Рассмотрим задачу декомпозиции смеси двух сигналов с меняющимся во времени периодом. Пусть общий вид сигналов в смеси задан в форме

$$f_t = \sum_{j=1}^N a_j \exp\left\{i \frac{2\pi}{G}(1 + \epsilon(t))jt\right\}, \quad (1.43)$$

где $\epsilon(t)$ — функция, значения которой случайным образом изменяются в моменты дискретного времени $t = nG$, $n=0, 1, \dots$ и подчинены распределению Гаусса с нулевым математическим ожиданием и дисперсией σ . Считая дисперсию распределения малой величиной такой, что наиболее вероятными являются

значения $|\epsilon(t)| << 1$, можно разложить экспоненты в (1.43) в ряд Тейлора по параметру $\epsilon(t)$.

Как это уже делалось ранее, приведем измерение спектра результата разложения в ряд Тейлора с произвольного номера отсчета τ с помощью ДПФ, ограниченного прямоугольным окном длительностью T , которая не кратна величине G

$$F_{q,\tau} = \frac{1}{T} \sum_{j=1}^N a_j \exp\left\{i \frac{2\pi}{T} \alpha j \tau\right\} \left(\frac{1 - W_j}{1 - R_{jq}} + i j \bar{\epsilon} \frac{2\pi}{G} \frac{R_{jq} + (TR_{jq} - T - 1)W_j}{[1 - R_{jq}]^2} \right), \quad (1.44)$$

где $\bar{\epsilon}$ — среднее значение функции на интервале спектрального окна T .

Безусловно, как и в предшествующем случае, наибольший вклад в q -ую компоненту спектра вносит компонента функции

с номером $j = \left[\frac{q}{\alpha} \right]$.

Поведение модуля коэффициента при компоненте $\left[\frac{q}{\alpha} \right]$ в

зависимости от q определяет минимальную разность между величинами основного тона сигналов, при которых возможно провести декомпозицию сигналов. Численная оценка показывает, что если в рамках одного спектрального окна отклонение основного тона компонент смеси равна 10 %, $T=512$, $G=100$, то при низких частотах ($q < 25$) это отклонение не влияет на ширину спектральных максимумов, и для того чтобы их разрешить, должно выполняться условие $[\alpha] - [\beta] \geq 2$, на более высоких частотах ($25 < q < 120$) для выделения спектральных максимумов должно выполняться условие $[\alpha] - [\beta] \geq 4$, и т.д. Рост ширины спектральных максимумов с ростом частоты определяется линейной зависимостью модуля коэффициента от частоты.

Метод декомпозиции компонент смеси сигналов с меняющимся периодом подобен методу, описанному ранее для смеси сигналов с постоянным периодом. Во-первых, необходимо най-

ти положения максимумов спектра, которые позволяют оценить округленные значения α и β . Во-вторых, поиск смещений спектральных максимумов от положений, предсказанных целочисленными значениями α и β , позволит найти поправки к этим значениям. Последняя задача состоит в оценке значения $\bar{\varepsilon}$.

Пусть в точке q найден спектральный максимум, который принадлежит определенной части смеси. Предполагая, что влияние другой части смеси на точку $q+1$ пренебрежимо мало, то для $\bar{\varepsilon}_k$ получим

$$\bar{\varepsilon}_q = \frac{F_{q+1} \frac{1-W_{\left[\frac{q}{\alpha}\right]}}{1-R_{\left[\frac{q}{\alpha}\right],q}} - F_q \frac{1-W_{\left[\frac{q}{\alpha}\right]}}{1-R_{\left[\frac{q}{\alpha}\right],q+1}}}{i \left[\frac{q}{\alpha} \right] \frac{2\pi}{G} (h_q - z_q)}, \quad (1.45)$$

$$\text{где } h_q = F_q \frac{R_{\left[\frac{q}{\alpha}\right],q+1} + \left(TR_{\left[\frac{q}{\alpha}\right],q+1} - T - 1 \right) W_{\left[\frac{q}{\alpha}\right]}}{\left[1 - R_{\left[\frac{q}{\alpha}\right],q+1} \right]^2},$$

$$z_q = F_{q+1} \frac{R_{\left[\frac{q}{\alpha}\right],q} + \left(TR_{\left[\frac{q}{\alpha}\right],q} - T - 1 \right) W_{\left[\frac{q}{\alpha}\right]}}{\left[1 - R_{\left[\frac{q}{\alpha}\right],q} \right]^2}.$$

Усреднения значения (1.45), найденные по всем спектральным максимумам данного элемента смеси, найдем оценку среднего отклонения от основной частоты сигнала.

Очевидно, что понятие среднего отклонения не дает информации о поведении основной частоты сигнала во времени, что приведет к потере точности выделенного сигнала относительно исходного элемента смеси.

Рассмотрим аддитивную смесь стационарного шума и периодического сигнала, причем в качестве шума будем использовать периодическую функцию с периодом много большим, чем период окна ДПФ. Для комплексного спектра этого сигнала справедливо

$$F_{q,\tau} = \frac{1}{T} \sum_{j=1}^N a_j \exp\left\{i \frac{2\pi}{T} \alpha j \tau\right\} \frac{1 - W_j(\alpha)}{1 - R_{jq}(\alpha)} + \frac{1}{T} \sum_{j=1}^M g_j \exp\left\{i \frac{2\pi}{T} \beta j \tau\right\} \frac{1 - W_j(\beta)}{1 - R_{jq}(\beta)}, \quad (1.46)$$

где $\alpha \gg 1$; $\beta \ll 1$; $\{a_j\}_N$; $\{g_j\}_M$ — спектральные компоненты речеподобного сигнала и шума, соответственно.

Поскольку параметр β мал, проведем разложение элементов последней суммы (1.46) в ряд Тейлора по этому параметру и учтем, что наибольший вклад в q -ую комплексную компоненту спектра вносит компонента речеподобного сигнала, равная округлению отношения q/α . Опуская громоздкие выкладки, получим выражение

$$F_{q,\tau} = \frac{1}{T} a_{\left[\frac{q}{\alpha}\right]} \exp\left\{i \frac{2\pi}{T} \left[\frac{q}{\alpha}\right] \alpha \tau\right\} \frac{1 - \exp\left\{i 2\pi \alpha \left[\frac{q}{\alpha}\right]\right\}}{1 - \exp\left\{i \frac{2\pi}{T} (\alpha \left[\frac{q}{\alpha}\right] - q)\right\}} + \sum_{j=1}^M g_j \left(1 + i \frac{2\pi}{T} j \beta \tau\right) \frac{\beta j}{\beta j - q}, \quad (1.47)$$

Заметим, что модуль функции $\frac{\beta j}{\beta j - q}$ убывает с отклонением j от величины $[q/\beta]$. Этот факт означает существование ограничения на значимые члены суммы, которые влияют на q -ую компоненту спектра. Пусть \bar{g}_k — среднее значение комплексных компонент шума, которые оказывают значимое влияние на q -ую компоненту измеряемого спектра, тогда формула (1.47) преобразуется к виду

$$F_{q,\tau} = \frac{1}{T} a_{\left[\frac{q}{\alpha}\right]} \exp\left\{i \frac{2\pi}{T} \left[\frac{q}{\alpha}\right] \alpha \tau\right\} \frac{1 - \exp\left\{i 2\pi \alpha \left[\frac{q}{\alpha}\right]\right\}}{1 - \exp\left\{i \frac{2\pi}{T} (\alpha \left[\frac{q}{\alpha}\right] - q)\right\}} + \\ + \bar{g}_q(\psi_q(\beta) + i \frac{2\pi}{T} \tau \varphi_q(\beta)),$$

где введены обозначения $\psi_q(\beta) = \beta \sum_{j=1}^M \frac{j}{\beta j - q}$, $\varphi_q(\beta) = \beta^2 \sum_{j=1}^M \frac{j^2}{\beta j - q}$.

Используем метод наименьших квадратов в форме

$$\sum_{\tau=1}^L \left(F_{q,\tau}^{(ex)} - \frac{1}{T} a_{\left[\frac{q}{\alpha}\right]} \exp\left\{i \frac{2\pi}{T} \left[\frac{q}{\alpha}\right] \alpha \tau\right\} \frac{1 - \exp\left\{i 2\pi \alpha \left[\frac{q}{\alpha}\right]\right\}}{1 - \exp\left\{i \frac{2\pi}{T} (\alpha \left[\frac{q}{\alpha}\right] - q)\right\}} + \right. \\ \left. + \bar{g}_q(\psi_q(\beta) + i \frac{2\pi}{T} \tau \varphi_q(\beta)) \right)^2 = 0,$$

где L — длительность последовательности точек, на которой определяется экспериментальный спектр $F_{q,\tau}^{(ex)}$. Дифференцируя последнее выражение по неизвестным параметрам α , β и \bar{g}_q , $a_{\left[\frac{q}{\alpha}\right]}$ получим нелинейную систему уравнений. Здесь мы не будем ее выписывать, лишь отметим, что она решается только численно.

В заключение этого раздела отметим, что перспектива развития данного направления имеет две стороны. Во-первых, необходимо искать теоретические решения задачи декомпозиции трех и более источников речеподобных сигналов. Во-вторых, естественные речевые сигналы имеют значительные вариации параметров, заключающие не только в изменениях основного тона,

которые рассматривались здесь, но и переходах из вокализованного в шипящий звук, а так же в значительных изменениях уровня огибающей сигнала.

1.2.2.3. Восстановление клипированного сигнала

Эффект клипирования возникает в цифровых телефонных каналах в случае, если динамический диапазон сигнала превышает динамический диапазон приемо-передающего устройства. Пример клипированного сигнала приведен на рисунке 1.8. Заметим, что клипирование сигнала приводит к искажению спектра звука.

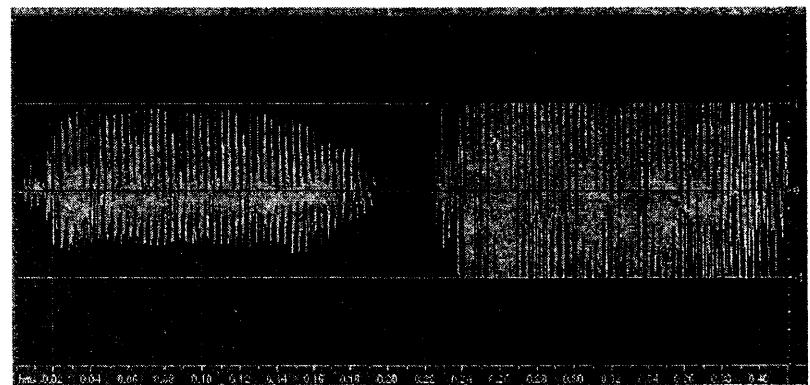


Рис. 1.8. Пример клипированного сигнала (справа) относительно исходного (слева)

Рассмотрим логарифмический спектр одного и того же звука — клипированного и неклипированного (рис. 1.9 и 1.10, соответственно). На рисунке 1.9 показан спектр звука «а», прошедшего ИКМ кодирование, а на рисунке 1.10 — спектр звука «а», прошедшего кодирование GSM и возникшим эффектом клипирования. Исследование проводилось с помощью программы Cool Edit Pro. Очевидно, что клипирование в значительной степени искажает спектр, сильно сглаживая (огрубляя) его.

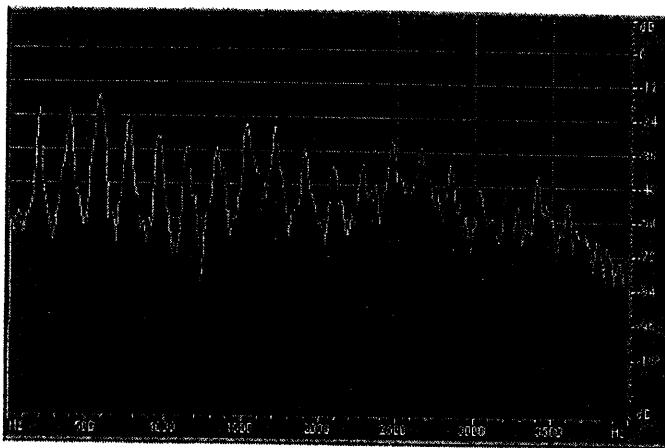


Рис. 1.9. Спектр речи при ИКМ кодировании

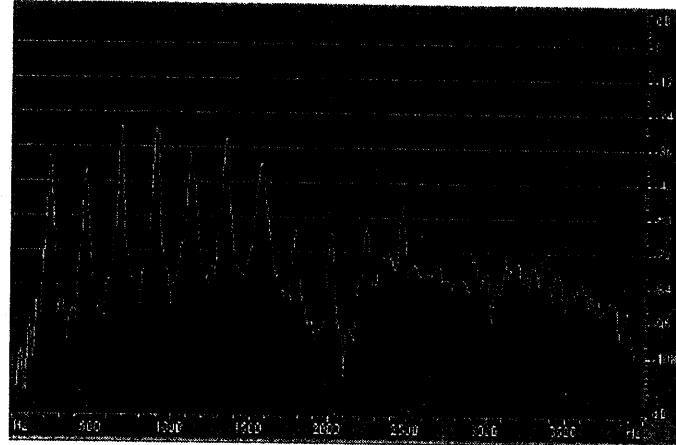


Рис. 1.10. Спектр речи при клиппировании сигнала

Практически часть спектра речи выше 2 кГц (рис. 1.10) представляет шумовую составляющую, не содержащую информацию о звуке. Сохраняется лишь общее распределение энергии по спектру. Очевидно, что это приводит к нарушениям работы систем обработки речи.

Рассмотрим интервал времени $[0, T]$ (здесь предполагается, что сигнал оцифрован через равные промежутки и время изме-

рено в количестве отсчетов), на котором произошел один акт клиппирования сигнала (рис. 1.11).

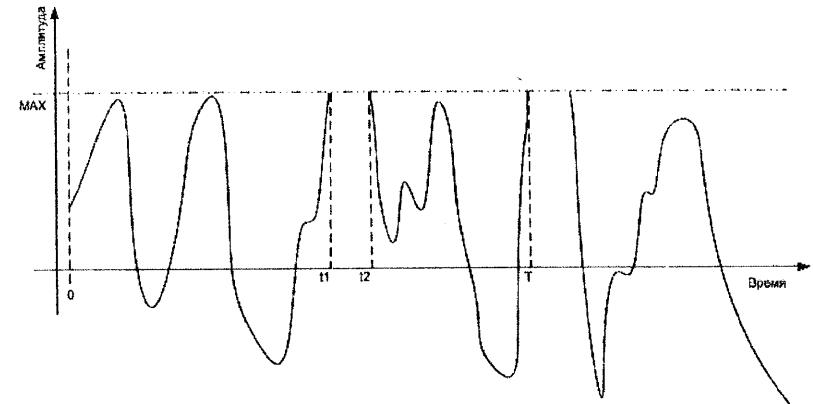


Рис. 1.11. Участок сигнала $[0, T]$, на котором произошел один акт клиппирования

Рассмотрим метод восстановления клиппированного участка с помощью интерполяции сигнала гармоническими функциями.

Пусть на интервалах $[0, t_1]$ и $[t_2, T]$ находятся M точек неискаженного сигнала. Интерполируем их с помощью конечной, взвешенной суммы комплексных экспонент с частотами кратными периоду T . Рассчитаем параметры этой суммы с помощью метода наименьших квадратов, т.е. минимизируем функционал

$$F = \sum_{t=0}^M \left(y_{m_t} - \sum_{j=0}^{[T/2]} a_j \exp \left\{ i \frac{2\pi}{T} j m_t \right\} \right)^2 \rightarrow \min. \quad (1.48)$$

Дифференцируя (1.48) по параметрам a_j получим систему линейных уравнений

$$\frac{\partial F}{\partial a_k} = \sum_{t=0}^M y_{m_t} \exp \left\{ -i \frac{2\pi}{T} k m_t \right\} - \sum_{j=0}^{[T/2]} B_{kj} a_j = 0,$$

$$\text{где } B_{kj} = \delta_{kj} - \sum_{t=t_1}^{t_2} \exp \left\{ i \frac{2\pi}{T} k t \right\} \exp \left\{ -i \frac{2\pi}{T} j t \right\} = \delta_{kj} - \frac{W_{k-j}^{t_0 \times M} (1 + W_{k-j}^{1 \times 2M})}{1 - W_{k-j}^1},$$

где δ_{kj} — символ Кронекера и введены обозначения

$$W'_{k,j} = \exp\left\{i\frac{2\pi}{T}(k-j)t\right\} \text{ и } t_2 = t_0 + \Delta t, \quad t_1 = t_0 - \Delta t.$$

Значения коэффициентов a_j позволяют с помощью суммирования ряда экспонент в точках интервала $[t_1, t_2]$ получить потерянные значения сигнала.

1.2.2.4. Восстановление потерянных блоков данных

Эффект потери пакетов данных при передаче возникает в телефонных сетях, использующих пакетный принцип передачи данных (например IP-телефония), и связан с перегрузкой сети, т.е., когда количество данных, которые необходимо передать, превышают пропускную способность сети. Пример принятого речевого сигнала при потере данных приведен на рисунке 1.12.

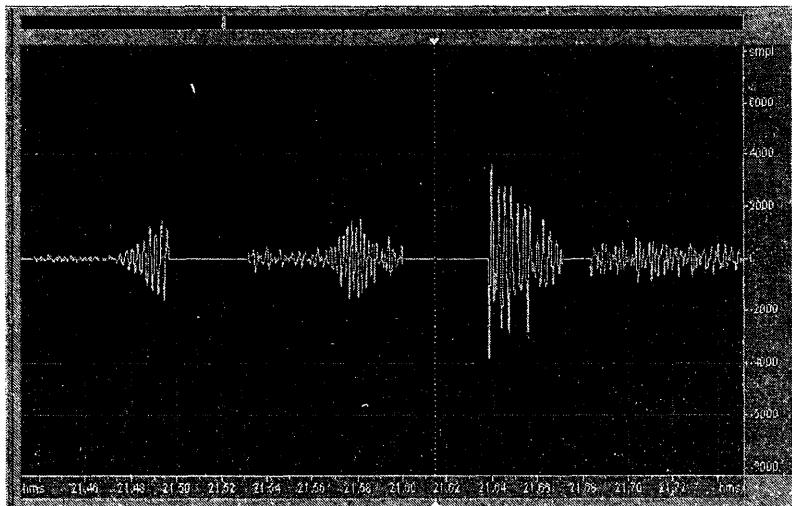


Рис. 1.12. Потеря блоков данных при пакетной передаче речи в телефонном канале

Строить процедуры восстановления потерянных блоков можно основываясь на двух основных идеях. Первая связана с

использованием системы распознавания фонем, которая должна искать наиболее вероятную потерянную фонему или цепочку фонем. На основании такой цепочки и их признаков, хранящихся в банке данных, формируется акустическая волна, окончательный вид которой определяется краевыми условиями, заданными параметрами сохранившихся блоков. Вторая идея более проста в реализации и основана на интерполяции параметров крайних целевых блоков, минуя этап распознавания фонем.

Здесь мы рассмотрим процедуру, основанную на втором подходе, при условии, что длительность потерянного блока не больше длительности минимального звучания фонемы. При оцифровке 8 кГц и передаче фреймами по 128 точек может быть потеряно не более пяти фреймов.

Для построения модели используем результаты приложения 3, где описана методика формирования формантных траекторий. Пример формантных траекторий проведен на рисунке 1.13. Здесь смешаны два графика: серым цветом выделена осцилограмма слова «самолет», а темные линии показывают поведение формант в этом слове.

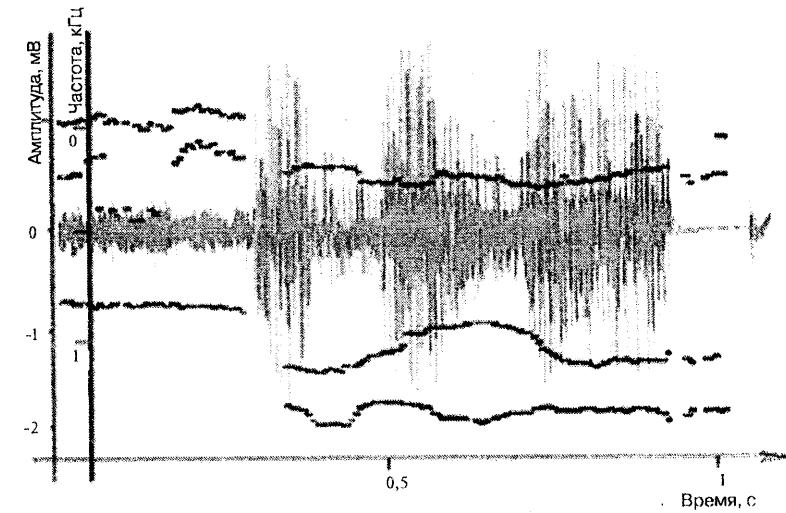


Рис. 1.13. Осцилограмма слова «самолет» и поведение формантных траекторий в этом слове

Пусть на интервале $[0, T]$ был потерян блок данных длительностью $[t_1, t_2]$, причем на промежутках времени $[0, t_1]$ и $[t_2, T]$ известны параметры формантных траекторий и поведения основного тона. Возникает вопрос, как траектории справа соответствуют траекториям слева. Строгие закономерности поведения траекторий неизвестны, поэтому можно воспользоваться линь вероятностной моделью.

Введем обозначения. Пусть слева имеется N_L формантных траекторий заданных своими конечными точками $\{\omega_{t_1}^L, A_{t_1}^L, \sigma_{1,2}^L\}$ и траектория частоты основного тона, заданная функцией Ω_t^L , а справа N_R формантных траекторий, заданных своими начальными точками $\{\omega_{t_2}^R, A_{t_2}^R, \sigma_{1,2}^R\}$ и траектория частоты основного тона, заданная функцией Ω_t^R .

Для построения вероятностной модели используем метод обучения, с помощью которого вычислим величины $p(\omega_i^R | \omega_j^L, M)$, которые являются условными вероятностями того, что i -ая траектория справа это продолжение j -ой траектории слева, M — число потерянных блоков $M = 1, \dots, 5$. Было бы желательно использовать в качестве аргумента условных вероятностей амплитуды и производные частоты траекторий, но это сделает решении задачи вычислительно более громоздким.

Для расчета условных вероятностей используем неискаженную выборку речи и будем считать количество попаданий траектории в компоненту спектра k — W_k^R , если M фреймов назад эта траектория проходила через компоненту спектра m — W_m^L , тогда

$$p(k | m) = \frac{W_k^R}{W_m^L}.$$

В дополнение к этой вероятности существуют вероятность затухания траектории прошедшей через m

$$p(-1 | m) = \frac{W_m^L - \sum_{k=1}^U W_k^R}{W_m^L},$$

где U — количество компонент, и вероятность возникновения новой траектории

$$p(k | -1) = \frac{W_k^R (-1)}{N},$$

где $W_k^R (-1)$ — количество прохождения новой траектории через компоненту k ; N — полное количество измерений.

Вероятность того, что N_L траекторий проходящих слева через точки $\{\omega_i^L\}$ пройдут через N_R точек $\{\omega_i^R\}$ справа с заданным соответствием равна

$$P(\{\omega_i^R\} | \{\omega_i^L\}; Y, M) = \prod_{i=1}^{N_L+1} p(\omega_{j_i}^R | \omega_i^L), \quad (1.49)$$

где Y — обозначение соответствия между точками; учтена возможность рождения и затухания траекторий на интервале $[t_1, t_2]$, т.е. $\omega_{N_R+1}^R = \omega_{N_L+1}^L = -1$.

Найдем такое соответствие между точками, при котором вероятность (1.49) достигает своего максимума. Причем будем считать, что разрешены только следующие перестановки индексов, если $\omega_i^L \rightarrow \omega_j^R$ и $\omega_k^L \rightarrow \omega_m^R$ и $\omega_i^L > \omega_k^L$, то $\omega_j^R > \omega_m^R$ (это требование равносильно запрету на пересечение траекторий), т.е.

$$Y^* = \arg \max_Y P(\{\omega_i^R\} | \{\omega_i^L\}; Y, M).$$

После того как соответствие между точками найдено, проведем интерполяцию траекторий, в том числе траектории частоты основного тона, дисперсий и амплитуд, любым из известных методов и фаз соответствующих компонент спектра.

Сумма гауссоид с интерполированными значениями амплитуд, дисперсий и частот формант определяют распределения амплитуд компонент спектра с кратностью, соответствующей

интерполированной частоте основного тона. Интерполированные значения фаз определяют фазу компонент. И в заключение с помощью ОДПФ синтезируем недостающий блок данных.

1.3. Выделение основного тона

Под выделением частоты ОТ обычно понимают определение как мгновенной частоты колебаний голосовых складок диктора [32], так и форму этих колебаний [7].

Мгновенная частота ОТ является значимым параметром практически во всех задачах классификации речи. В задачах идентификации дикторов по значению средней частоты ОТ говорящего возможно создать, по крайней мере, два хорошо различимых класса дикторов (мужчины и женщины). В задачах распознавания речи по наличию ОТ, ее можно классифицировать на вокализованные (гласные и звонкие согласные) и невокализованные звуки (шипящие и глухие согласные). В задачах классификации эмоциональных состояний говорящего, используется динамика частоты ОТ и различных показателей формы импульса ОТ (длительность переднего фронта, показатель асимметрии [7] и т.п.).

С момента появления первых работ посвященных методам выделения ОТ [33, 34], они претерпели существенные изменения и развитие. Здесь остановимся на тех методах, которые нам кажутся наиболее интересными.

Рассмотрим *автокорреляционный метод* выделения ОТ [32] в цифровой форме

$$B_{\tau} = \vartheta_{\tau} \sum_{k=t-C}^{t-1} \gamma_{t-k} y_k y_{k-\tau}, \quad (1.50)$$

где ϑ_{τ} , γ_t — весовые функции; y_t — речевой сигнал; величина C задает интервал суммирования.

На рисунке 1.14 на графиках 2—4 показано поведение во времени автокорреляционной функции (1.50) при различных

значениях смещения: график (2) $\tau = 3$ мс, график 3 $\tau = 6$ мс, график 4 $\tau = 10$ мс, при условиях:

$$\gamma_t = \exp\{-\alpha t\} \text{ и } \vartheta_{\tau} = 1, \forall \tau. \quad (1.51)$$

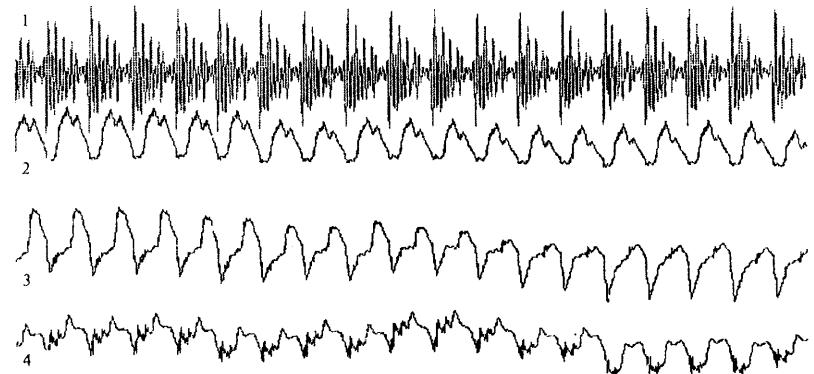


Рис. 1.14. Осцилограмма речи и поведение во времени автокорреляционных функций при различных значениях параметра смещения

Рассчитаем АЧХ $H(\omega, t, \tau)$ с помощью метода предложенного в [35] при выполнении условия (1.51). Подставим в (1.50) в качестве входной функции f , комплексную экспоненту $\exp\{-i\omega t\}$, тогда

$$H(\omega, t, \tau) = \operatorname{Re} \left(e^{-i\omega t} \sum_{k=t-C}^{t-1} e^{-\alpha(t-k)} e^{i\omega k} e^{i\omega(t-k-\tau)} \right) = \operatorname{Re} \left(e^{-i\omega(t-\tau)} \frac{\beta e^{-2i\omega C} - 1}{1 - \rho e^{2i\omega}} \right), \quad (1.52)$$

где $\beta = \exp\{-\alpha C\}$, $\rho = \exp(\alpha)$.

Результаты численного расчета (1.52) при условии $C = 20$ мс, $\alpha = 0,05$ показаны на рисунке 1.15.

Этот результат означает, что преобразование (1.50) представляет собой низкочастотный фильтр. В зависимости от разности момента времени t и величины смещения τ реализуется тот или иной срез АЧХ.

Дальнейшая обработка автокорреляционной функции B_{τ} состоит в том, выбираются несколько равноудаленных значений величины $\tau = n\Delta\tau$, $n = 1, \dots, Q$ и B_{τ} фильтруются полосовыми фильтрами с полосой пропускания 80—450 Гц (описаны кла-

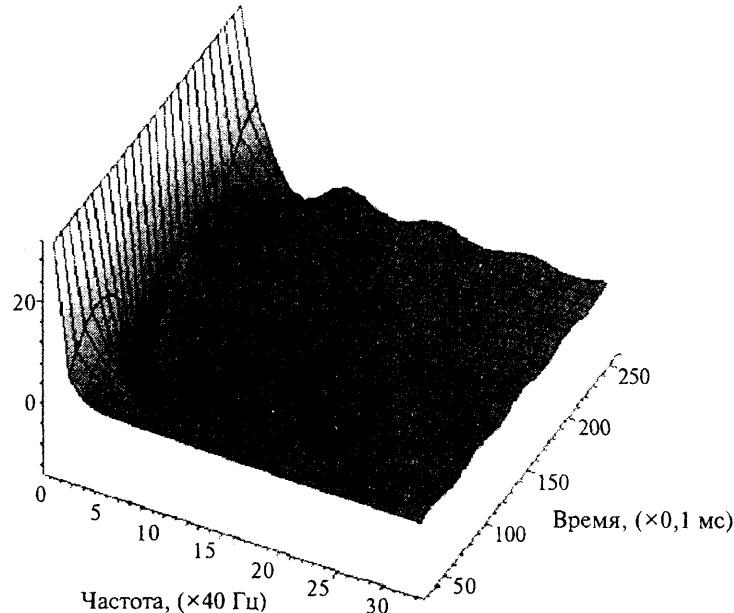


Рис. 1.15. АЧХ автокорреляционного преобразования

сов данных фильтров можно найти в монографии [35]). Выбор такой полосы пропускания связан с эмпирическими данными о диапазоне частоты ОТ.

В каждом из отфильтрованных сигналов производится поиск максимумов. Предполагается, что в каждом сигнале максимумы, определенные составляющими с частотой больше частоты ОТ случайны, а максимумы, определенные составляющей ОТ, регулярны. На основании этого предположения, достаточно следить за синхронным появлением максимумов для всех отфильтрованных сигналов (Q штук), чтобы выделить составляющую ОТ. По значению скважности между синхронно поступающими максимумами определяется мгновенная частота ОТ.

Другим классом методов выделения основного тона являются *пиковые методы*, модификации которых описаны во многих трудах [36,37]. В качестве примера, показывающего принципы

работы методов этого типа, используем систему, описанную в работе [38].

Алгоритм состоит из четырех операций:

- 1) полосовая фильтрация речевого сигнала;
- 2) образование шести функций по экстремумам отфильтрованного сигнала;
- 3) получение на основе значений вышеупомянутых функций шести оценок ОТ в шести одинаковых измерителях;
- 4) принятие окончательного решения на основе оценок элементарных измерителей ОТ.

Основное назначение полосового фильтра состоит в подавлении гармоник априорно не имеющих отношения к гармоникам основного тона. Как уже было сказано выше, для решения этой задачи используются фильтры с полосой пропускания 80–450 Гц.

При выполнении второй операции для всех экстремумов отфильтрованного колебания формируются импульсы различной амплитуды. Для каждого максимума импульсы с амплитудами m_1, m_2, m_3 , а для каждого минимума импульсы с амплитудой m_4, m_5, m_6 , как показано на рисунке 1.16

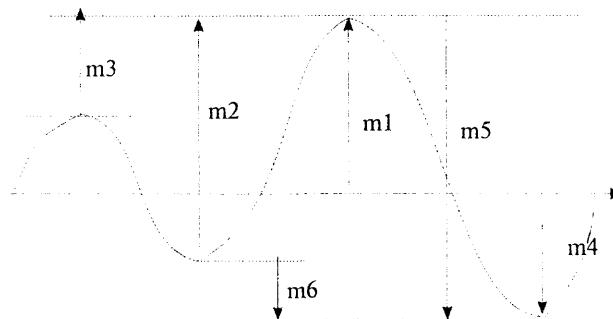


Рис. 1.16. Метод построения различных пиков при анализе волны

Шесть импульсных последовательностей поступают на входы шести одинаковых измерителей периода. По существу каждый измеритель является пиковым детектором с управляемой цепью разряда. После приема каждого импульса следует интер-

вал запирания (в течение которого детектор не принимает импульсов), а за ним — интервал экспоненциального разряда детектора. Если на этом втором интервале приходит импульс, превышающий напряжение в цепи разряда, то он детектируется, и процесс запирания и разряда повторяется.

Окончательный период ОТ определяется по величине наиболее вероятного совпадения оценок линий каждого из шести детекторов. При принятии решения о совпадении двух оценок представляется более целесообразным рассматривать их отношение, а не разность. Часто последовательные измерения заметно отличаются, поэтому полезно ввести несколько пороговых величин для определения совпадений, и при вычислении оценки периода выбирать ту из них, которая дает наиболее разумный ответ.

Опишем метод, основанный на *повторном анализе исходного сигнала*.

На первом этапе анализа производится предварительная оценка частоты ОТ [39]. Известно, что спектр Фурье вокализованного звука при условии, что длительность окна анализа больше двух периодов ОТ, имеет выраженный линейчатый характер (рис. 1.17),

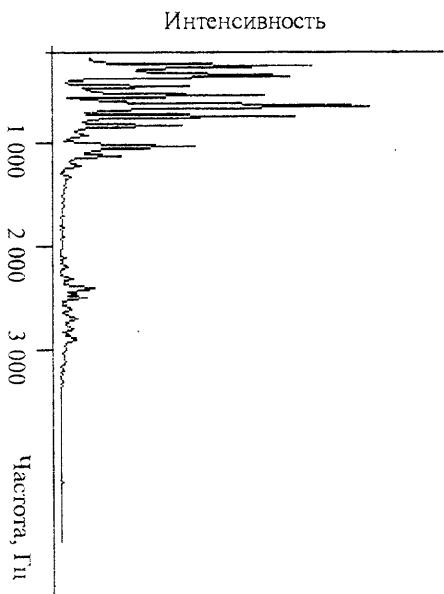


Рис. 1.17. Пример линейчатого спектра звука «a»

причем расстояния между линиями спектра распределены около частоты ОТ. Таким образом, по расстояниям между линиями спектра возможно оценить частоту ОТ.

Для вычисления оценки частоты основного тона на интервале T анализа спектра Фурье, взятого в момент времени t , проинтегрируем произведение нормированного спектра

$$\tilde{F}_j = \frac{h_j}{\sqrt{\sum_{i=0}^T h_i^2}},$$

и сумму параметрических функций, полученных при дифференциировании функций Гаусса и зависящих от параметра t , в соответствии с формулой [39]:

$$\gamma_j(\tau) = \sum_{i=0}^T \tilde{F}_{ij} \sum_{l=1}^n (1 - \alpha(j - \tau i)^2) \exp \left\{ -\alpha(j - \tau i)^2 \right\}, \quad (1.53)$$

где τ — величина смещения между двумя соседними функциями (она собственно и определяет частоту основного тона); $n = \left[\frac{T}{\tau} \right] -$ целое количество функций, уменьшающиеся в диапазоне спектра, α — параметр модели равный $\frac{32}{\tau^2}$.

Если окно ДПФ покрывает участок невокализованного звука, то величина $\max_{\tau} \gamma_j(\tau)$ будет близка к нулю, поскольку спектр такого звука рассредоточен по всему диапазону частот. Этот факт дает возможность установить порог Q (его эмпирическое значение равно 0,2) на значение $\max_{\tau} \gamma_j(\tau)$ для выделения вокализованных участков. Аргумент максимума интеграла (1.53) при усlovии $\max_{\tau} \gamma_j(\tau) > Q$ показывает оценку значения частоты ОТ, т.е. $\omega_{OT,j} = \arg \max_{\tau} (\gamma_j(\tau))$.

Очевидно, что на каждом интервале анализа спектра Фурье

вокализованных звуков во фразе будет получено свое значение ОТ. Этот эффект будет связан с двумя факторами: а) изменением частоты основного тона со временем; б) точностью измерения спектра.

Для получения средней оценочной частоты основного тона усредним полученные оценки по всем интервалам

$$\bar{\omega}_{OT} = \frac{1}{D} \sum_{i=1}^D \omega_{OT,i},$$

где D — количество интервалов анализа.

Второй этап состоит в том, что выбирается некоторый низкочастотный фильтр с частотой среза, равной найденной оценке частоты ОТ, и исходный сигнал пропускается сквозь этот фильтр. Для этих целей ранее авторами использовался фильтр Баттерворта 5-го порядка [40]. В выходном сигнале фильтра ведется поиск максимумов.

Пусть моменты времени появления максимумов образуют последовательность $\{t_1, t_2, \dots, t_M\}$. Разность между двумя соседними моментами времени $L_i = t_i - t_{i-1}$ означает длительность импульса основного тона. Психоакустические данные показывают, что изменение длительности импульсов основного тона в слоге (см. например, [41]) можно аппроксимировать квадратичной функцией. Используем этот факт.

Выберем в последовательности $\{L_i\}$ подпоследовательность $\{L_i\}$, в которой i изменяется от некоторого k до $k+R$, где R длина этой подпоследовательности.

Минимизируем сумму квадратов разностей

$$F_k(a, b, c) = \sum_{i=k}^{k+R} (L_i - (a_k(i-k)^2 + b_k(i-k) + c_k))^2, \quad (1.54)$$

и найдем коэффициенты a_k, b_k, c_k .

Все элементы подпоследовательности являются интервалами основного тона, если ошибка аппроксимации удовлетворяет условию

$$F_k(a, b, c) < Q_F, \quad (1.55)$$

где Q_F — параметр модели.

Если же для данной подпоследовательности условие (1.55) не выполняется, то считается, что элемент подпоследовательности L_k не несет информации об ОТ. В дальнейшем выбирается новая подпоследовательность той же длины R , смещенная относительно предыдущей подпоследовательности на единицу. Если посредством J_n обозначить такие L_i , которые удовлетворяют условию (1.55), то речь становится сегментированной именно интервалами J_n .

Необходимо отметить, что если требуется повышенная точность определения мгновенной частоты основного тона и не предъявляются высокие требования на скорость обработки сигнала, то можно использовать одновременную обработку сигнала несколькими выделителями ОТ.

Функциональное устройство алгоритма, включающего в себя три выделителя ОТ, показано на рисунке 1.18.

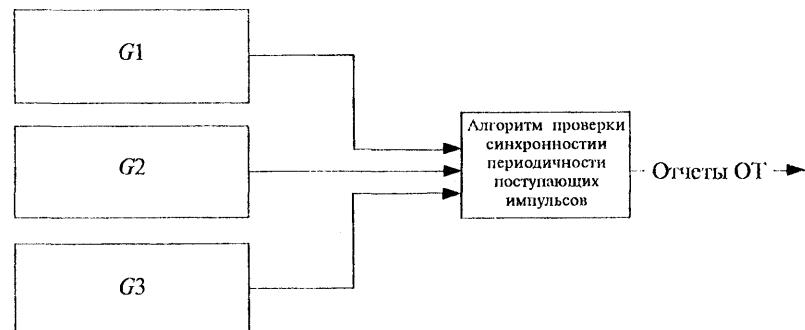


Рис. 1.18. Функциональная схема устройства определения мгновенной частоты основного тона

Символами $G1, G2, G3$ обозначены различные методы выделения ОТ.

Для примера, в качестве первого метода выделения ОТ ($G1$) выберем автокорреляционный метод (1.50). Найдем максимумы отфильтрованной автокорреляционной функции (1.50). Фильтрацию лучше всего производить фильтром низких частот с час-

тотой среза 400 Гц. Выходную функцию алгоритма G_1 , записем в виде суммы дельта-функций Дирака

$$g_1(t) = \sum_k \delta(t - t_k), \quad (1.56)$$

где t_k — время появления k -го максимума. В качестве второго метода (G_2) выберем уже описанный пиковый метод, выходную функцию которого тоже запишем в виде (1.56) и обозначим $g_2(t)$. Третий метод (G_3) связан с тем, что поведение ОТ аппроксимируется экспоненциальной функцией $d(t) = b \exp(-\beta(t - \chi))$. Величина b принимает значение $|y_i|$, а $\chi = t$ в случае $d(t) \leq |y_i|$. Если пронумеровать моменты времени χ , в которые величина b изменяет свое значение, то результатом работы алгоритма станет функция аналогичная (1.56), обозначим ее $g_3(t)$.

Окончательное решение о значении мгновенной частоты ОТ или его отсутствии принимается на основе поведения решения дифференциального уравнения первого порядка

$$\frac{d\phi}{dt} + \gamma\phi = \sum_{i=1}^3 g_i(t),$$

с начальным условием $\phi(0) = 0$, которое проверяется на выполнение порогового условия

$$\phi(t) = Q, \quad (1.57)$$

где Q — порог, принимающий значения в интервале от 1,5 до 1,7. Моменты времени, в которые выполняется условие (1.57), образуют последовательность $\{t_i\}$. Разность двух следующих друг

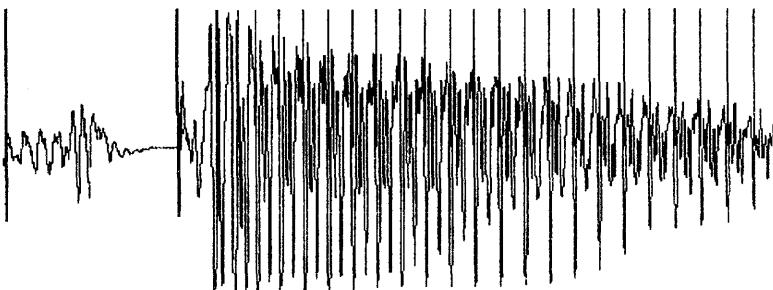


Рис. 1.19. Интервалы ОТ найденные при одновременной обработке сигнала несколькими алгоритмами

за другом моментов $L_i = t_{i+1} - t_i$ образуют новую последовательность L , которая аппроксимируется с помощью полинома второго порядка (1.54), и при выполнении условия (1.55) принимается решение о том, является ли интервал L_i основным тоном или нет. На рисунке 1.19 показаны сегменты, соответствующие интервалам ОТ.

Сравнение точности работы описанного алгоритма выделения основного тона с алгоритмом, основанном на вычислении автокорреляционной функции (1.50), показывает, что если ошибка первого алгоритма составляет 4–9 % в зависимости от высоты голоса, то у второго достигает 12–19 %.

1.4. Параметризация и нормализация речевого сигнала. Оценка значимости спектральных компонент

Роль процедуры параметризации выделить наиболее информативные параметры речи. В зависимости от типа задачи классификации может подходить тот или иной вид параметризации (рис. 2). Например, основное требование к виду параметризации при дикторонезависимом распознавании речи в том, чтобы она как можно сильнее «слаживала» индивидуальные особенности голосов дикторов, и обратная задача должна решаться при параметризации в системе текстонезависимой идентификации дикторов.

Ключевым при параметризации является предположение о стационарности речевого сигнала на промежутках времени порядка нескольких миллисекунд. Таким образом, в ходе анализа речь разбивается на блоки данных. Обычно такие блоки называются окнами. На основе данных окна вычисляется вектор признаков, который является основой для решения любой из задач обработки речи.

Один из методов параметризации описан в приложении 3.1 и связан с вычислением формантных наборов. Здесь мы опишем еще два наиболее распространенных метода, которые основаны на психоакустических данных о свойствах слуха человека.

В ходе психоакустических исследований было выяснено два факта:

- 1) субъективная интенсивность воспринимаемого звука существенно зависит от частоты этого звука;
- 2) субъективная частота воспринимаемого звука отлична от его физической частоты.

Субъективную громкость звука принято описывать с помощью функций равной громкости [13]. Физическая суть этих функций состоит в определении уровня звукового давления тонального сигнала заданной частоты, при условии совпадения его субъективной громкости с субъективной громкостью тонального сигнала частоты 1 000 Гц с заданным давлением (рис. 1.20).

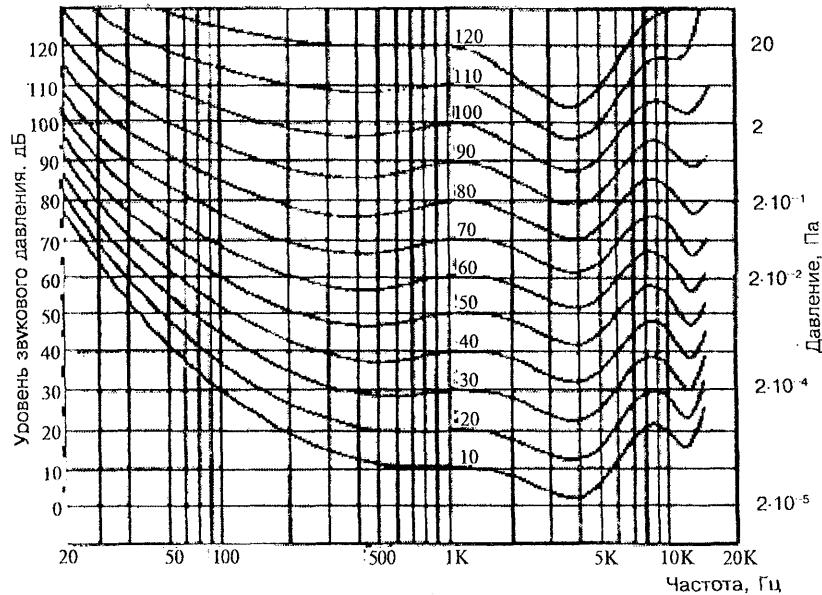


Рис. 1.20. Функции равной громкости

Характер субъективного восприятия частоты тонального сигнала принято описывать масштабной шкалой, которая названа мел шкалой (рис. 1.21).

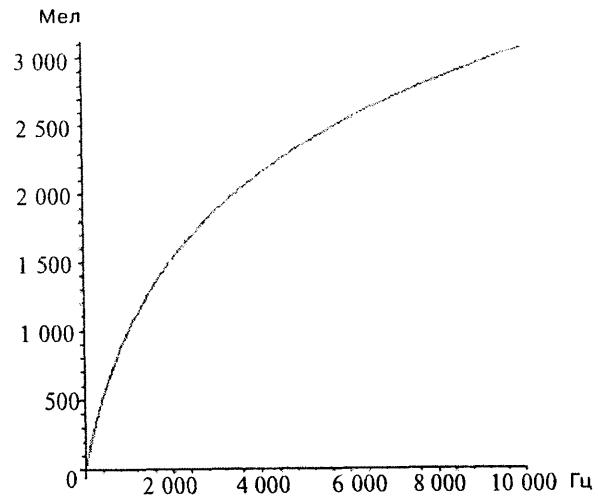


Рис. 1.21. Мел шкала

Она показывает, что ощущение частоты тонального сигнала изменяется по логарифмическому закону

$$Mel(f) = 2595 \lg \left(1 + \frac{f}{700} \right), \quad (1.58)$$

который был введен эмпирически.

1.4.1. Вычисление логарифмически масштабированных кепстральных коэффициентов

На рисунке 1.22. представлен процесс получения MFCC.

На первом шаге вычисляются компоненты спектра Фурье. Далее спектр сглаживается при помощи операции свертки с треугольным окном, расположение которых подчиняется закону (1.52), как показано на рисунке 1.23.

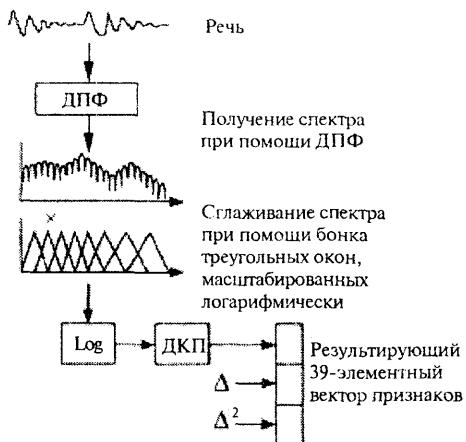


Рис. 1.22. Структура вычисления кепстральных коэффициентов

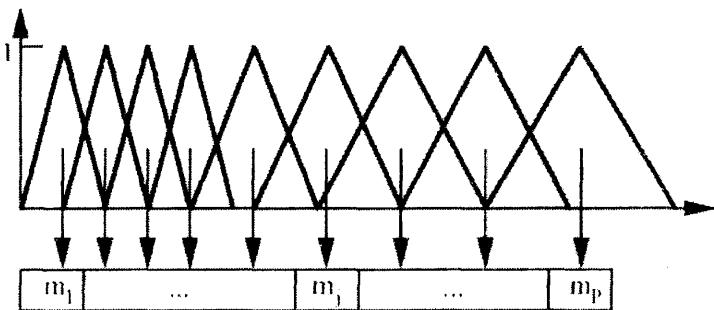


Рис. 1.23. Логарифмическое распределение окон для сглаживания спектра

Формально операция сглаживания определяется следующим образом:

$$M_j = \sum_{i=0}^{\frac{N}{2}} m_{ji} c_i, j = 0, 1, \dots, K, \quad (1.59)$$

где $c_i, i=0, \dots, \frac{N}{2}$ — спектр сигнала; K — количество треугольных окон; m_{ji} — величина j -ой оконной функции. Оконная функция

выражается через величины a_j — начало j -го треугольного окна — следующим образом:

$$m_{ji} = \begin{cases} \frac{i - a_{j-1}}{a_j - a_{j-1}}, & \text{если } a_{j-1} \leq i \leq a_j, \\ \frac{i - a_{j+1}}{a_j - a_{j+1}}, & \text{если } a_j < i \leq a_{j+1}, \\ 0 & \text{в другом случае} \end{cases} \quad (1.60)$$

Окончательно MFCC-коэффициенты получаются при помощи дискретного косинус-преобразования:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N M_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right).$$

MFCC-преобразование является по существу сжатием спектральных данных, что значительно упрощает моделирование речи за счет снижения размерности вектора признаков.

1.4.2. Вычисление масштабированных коэффициентов линейного предсказания

Процесс вычисления мел коэффициентов линейного предсказания приведен на рисунке 1.24.

На пути вычисления выполняются следующие шаги:

- 1) вычисление спектра Фурье по N отсчетам сигнала s_1, \dots, s_N ;
- 2) сглаживание спектра Фурье при помощи операции свертки по перекрывающимся треугольным окнам в соответствии с (1.59) — (1.60);
- 3) умножение полученных коэффициентов на функцию равной громкости:

$$M'_j = M_j E(\omega_j),$$

которая задана эмпирической формулой

$$E(\omega) = \frac{(\omega^2 + 1200^2)\omega^4}{(\omega^2 + 400^2)^2(\omega^2 + 3100^2)},$$

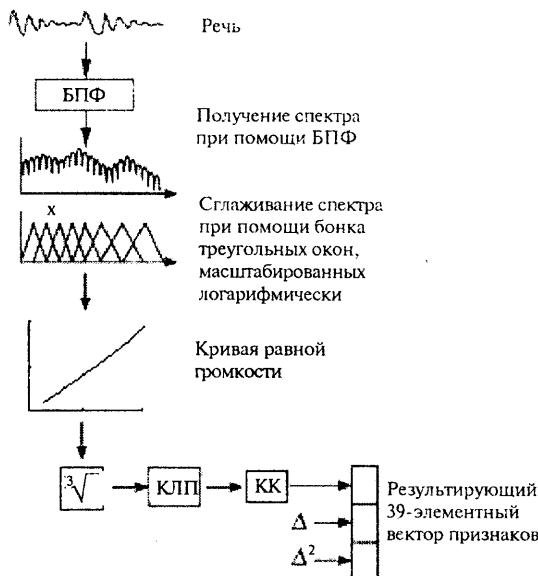


Рис. 1.24. Схема вычисления мел коэффициентов линейного предсказания

где ω_j — частота, соответствующая центру j -го треугольного окна в шаге 2. Этот шаг позволяет учесть зависимость чувствительности человеческого слуха от частоты;

4) извлечение кубического корня: $M''_j = \sqrt[3]{M'_j}$;

5) расчет обратного преобразования Фурье на основе значений M''_j ;

6) вычисление коэффициентов линейного предсказания вновь образованного, сглаженного сигнала.

Остановимся на методе расчета коэффициентов линейного предсказания. Передаточная функция голосового тракта моделируется передаточной функцией идеального фильтра:

$$H(z) = \frac{1}{\sum_{i=0}^p b_i z^{-i}},$$

где p — число полюсов, $b_0 \equiv 1$. Коэффициенты фильтра b выбираются так, чтобы минимизировать среднеквадратическую ошибку предсказания по всему окну анализа.

Пусть имеется последовательность отсчетов s_n , $n = 1 \dots N$. Первые $p+1$ элементов автокорреляционной последовательности рассчитываются по формуле

$$r_i = \sum_{j=1}^{N-i} s_j s_{j+i}, \quad i = 0, \dots, p.$$

Коэффициенты фильтра b_i рассчитываются рекурсивно с использованием вспомогательных коэффициентов k_p , которые можно интерпретировать как коэффициенты отражения эквивалентного акустического волновода и ошибки предсказания E , начальное значение которой равно r_0 . Пусть $\{k_j^{(i-1)}\}$ и $\{b_j^{(i-1)}\}$ — коэффициенты отражения и коэффициенты фильтра порядка $i-1$. Тогда фильтр порядка i может быть пересчитан в три шага. Сначала пересчитывается новый набор коэффициентов отражения:

$$k_j^{(i)} = k_j^{(i-1)} \text{ для } j = 1, \dots, i-1,$$

и

$$k_i^{(i)} = \frac{\left\{ r_i + \sum_{j=1}^{i-1} b_j^{(i-1)} r_{i-j} \right\}}{E^{(i-1)}}.$$

Далее вычисляется ошибка предсказания

$$E^{(i)} = (1 - k_i^{(i)} k_i^{(i)}) E^{(i-1)}.$$

И окончательно рассчитываются новые коэффициенты фильтра:

$$b_j^{(i)} = b_j^{(i-1)} - k_i^{(i)} b_{i-j}^{(i-1)} \text{ для } j = 1, \dots, i-1$$

и

$$b_i^{(i)} = -k_i^{(i)}.$$

Этот процесс повторяется от $i=1$ до требуемого порядка фильтра $i=p$.

В нашем случае в качестве входной последовательности отсчетов выступают отсчеты сглаженного, синтезированного сигнала, полученные на этапе 5.

7) Расчет кепстральных коэффициентов на основе коэффициентов линейного предсказания. Известно, что, основываясь на коэффициентах линейного предсказания, можно рассчитать спектр. Поскольку в нашем случае на этапе 6 рассчитывались коэффициенты линейного предсказания от спектра, то, пересчитывая через них спектр, получим кепстр. Для этих целей существует простая рекурсивная формула:

$$c_n = -b_n + \frac{1}{n} \sum_{i=1}^{n-1} (n-i)b_i c_{n-i}.$$

Для таких способов спектральных оценок как вычисление мел кепстральных коэффициентов или вычисление мел коэффициентов линейного предсказания часто применяют фильтр RASTA. Эту операцию можно классифицировать, как сглаживание в соответствии с крайней правой колонкой рисунка 2.

Метод фильтрации RASTA основан на предположении, что поведение компонент спектральных оценок речи — это медленно меняющиеся функции, которые не могут испытывать резких скачков, поэтому для каждой компоненты можно ввести фильтрацию, сглаживающую нежелательные колебания шумовой природы.

Для этих целей часто используют низкочастотный фильтр с передаточной характеристикой

$$H(z) = 0,1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0,98z^{-1}},$$

одинаковой для каждой компоненты.

Здесь может показаться, что применение одинаковых фильтров для каждой компоненты в соответствии с вычислениями (1.29) равносильно использованию некоторого единого фильтра во временной области. Это не так, поскольку фильтр RASTA используют для функций квадратов компонент ДПФ, а утверждение (1.29) доказывалось для компонент ДПФ.

В дополнении к кепстральным коэффициентам или коэффициентам линейного предсказания вектор признаков расширяют их первыми и вторыми производными, которые рассчитывают по известным разностным формулам

$$d_i = \frac{1}{\Theta} (\Delta a_i - \frac{1}{2} \Delta^2 a_i + \frac{1}{3} \Delta^3 a_i - \dots);$$

$$g_i = \frac{1}{\Theta^2} (\Delta^2 a_i - \Delta^3 a_i + \frac{11}{12} \Delta^4 a_i - \frac{5}{6} \Delta^5 a_i - \dots),$$

где $\Delta a_i = a_i - a_{i-1}$; Θ — шаг анализа коэффициентов. Количество коэффициентов для расчета производных выбирается в зависимости от соотношения шага окна анализа и минимальной длительности фонемы. Эти соображения нами были описаны ранее.

1.4.3. Селекция значимых компонент спектральной оценки

Какой бы путь оценки спектральных характеристик речевого сигнала не был выбран, очень важно иметь представление о том, какие из компонент спектра являются значимыми, и какие функциональные зависимости между этими компонентами существуют, которые позволяли бы наилучшим образом отличать одно состояние сигнала от другого. Например, можно выяснить значимые компоненты и функциональные зависимости между ними для различия какой-либо пары фонем, или какой-либо пары дикторов и т.д. Такую возможность предоставляет МГУА [23].

Далее будем рассуждать на примере фонем. Пусть для F фонем даны их акустические реализации, которые состоят из k_f векторов наблюдений $\mathbf{h}_f^{(0)}$, где f — индекс фонемы в алфавите.

В соответствии с МГУА, введем преобразование, которое отображает в двухмерном пространстве i -ую и j -ую компоненты n -го вектора наблюдений f -ой фонемы на некоторую поверхность второго порядка

$$\Phi(\mathbf{h}_n^{(f)}) = a_{1ij} h_{ni}^{(f)} h_{nj}^{(f)} + a_{2ij} (h_{ni}^{(f)})^2 + a_{3ij} (h_{nj}^{(f)})^2. \quad (1.61)$$

Аналогичное преобразование (1.61) задает отображение i -ой и j -ой компонент фонемы g . Для вычисления параметров $a_{ij} - a_{3ij}$ в (1.61) необходимо задаться некоторым критерием, на основе которого можно наилучшим образом разделять области компонент вектора наблюдений фонем g и f . На рисунке 1.25, для примера, приведены некоторые области компонент векторов признаков А и В, и задана некоторая поверхность второго порядка.

Для поиска значений параметров преобразования (1.61) относительно каждой пары фонем (всего C_F^2 пар фонем) минимизируем функционал

$$\Delta_{ij}^{(fg)} = - \left(\sum_{n=0}^{k_f-1} r^2(\bar{\mathbf{h}}^{(f)}, G_n^{(f)}) + \sum_{m=0}^{k_g-1} r^2(\bar{\mathbf{h}}^{(g)}, G_m^{(g)}) \right) = \min, \quad (1.62)$$

где $r(x_1, x_2)$ — евклидово расстояние между точками x_1, x_2 ; $\bar{\mathbf{h}}^{(w)}$ — вектор средних значений с компонентами

$$\bar{\mathbf{h}}^{(w)} = \frac{1}{k_w} \sum_{n=0}^{k_w-1} \mathbf{h}_n^{(w)};$$

$G_n^{(w)}$ — точка пресечения линии, проведенной из точки $\bar{\mathbf{h}}^{(w)}$ и проходящей через точку $\mathbf{h}_n^{(w)}$ с ближайшей точкой линии удовлетворяющей условию $\phi(\mathbf{h}_n^{(w)}) = 0$.

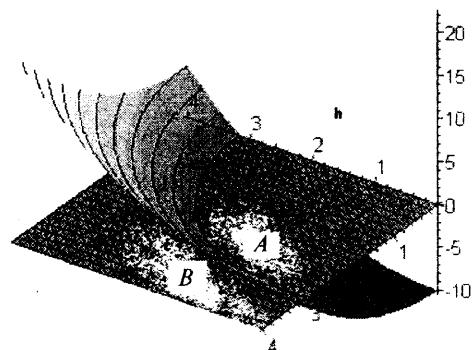


Рис. 1.25. Расположение областей компонент фонем относительно поверхности второго порядка

Использование такого вида функционала можно пояснить с помощью графика, представленного на рисунке 1.26, где области А и В соответствуют парам компонент фонем f и g , G — линия пересечения поверхности второго порядка с плоскостью (h_i, h_j) .

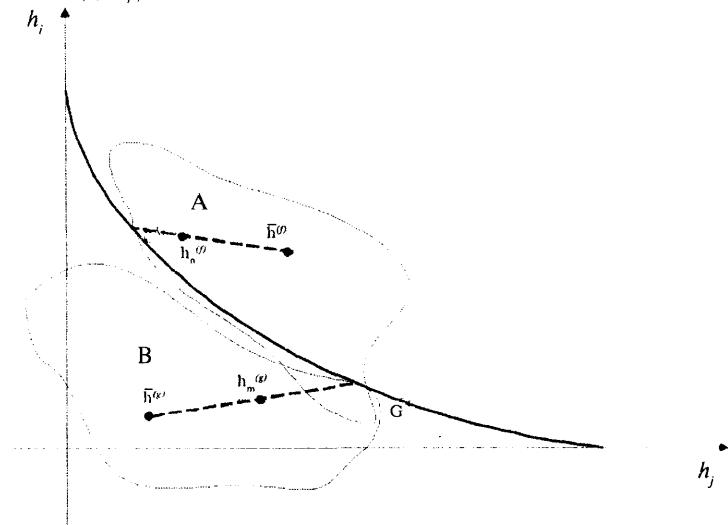


Рис. 1.26. Способ вычисления расстояний до линии пересечения поверхности второго порядка с плоскостью (h_i, h_j)

Поскольку граница G разделяет пространство параметров на две области, то квадрат расстояния от средней точки $\bar{\mathbf{h}}^{(w)}$ до некоторой n -ой точки наблюдения этой фонемы $\mathbf{h}_n^{(w)} - r^2(\bar{\mathbf{h}}^{(w)}, \mathbf{h}_n^{(w)})$ должен быть меньшим, чем квадрат расстояния от средней точки до границы G , отсчитанное в направлении точки $\mathbf{h}_n^{(w)} - r^2(\bar{\mathbf{h}}^{(w)}, G_n^{(w)})$ (рис. 1.26). Некоторые из точек $\mathbf{h}_n^{(w)}$ могут лежать по другую сторону границы относительно средней точки $\bar{\mathbf{h}}^{(w)}$, и тогда это условие нарушается, т.е. в этом случае $r^2(\bar{\mathbf{h}}^{(w)}, \mathbf{h}_n^{(w)}) > r^2(\bar{\mathbf{h}}^{(w)}, G_n^{(w)})$. Разность

$\varepsilon = r^2(\bar{\mathbf{h}}^{(w)}, \mathbf{h}_n^{(w)}) - r^2(\bar{\mathbf{h}}^{(w)}, G_n^{(w)})$, которая образует функционал

$$\Delta_{ij}^{(fg)} = \sum_{n=0}^{k_f-1} r^2(\bar{\mathbf{h}}^{(f)}, \mathbf{h}_n^{(f)}) - r^2(\bar{\mathbf{h}}^{(f)}, G_n^{(f)}) + \sum_{m=0}^{k_g-1} r^2(\bar{\mathbf{h}}^{(g)}, \mathbf{h}_m^{(g)}) - r^2(\bar{\mathbf{h}}^{(g)}, G_m^{(g)}), \quad (1.63)$$

можно использовать в качестве оценки качества сепарации точек наблюдений. Поскольку при выполнении операции минимизации, дифференцирование функционала (1.63) происходит по параметрам $a_{1j} - a_{3j}$, определяющим границу, то членами функционала вида $r^2(\bar{\mathbf{h}}^{(w)}, \mathbf{h}_n^{(w)})$ можно пренебречь и получить функционал в виде (1.62).

Запишем функционал (1.62) в явном виде

$$\Delta_{ij}^{(fg)} = - \left(\sum_{n=0}^{k_f-1} (\bar{h}_i^{(f)} - x_n^{(f)})^2 + (\bar{h}_j^{(f)} - \lambda_n^{(f)} x_n^{(f)} - p_n^{(f)})^2 + \sum_{n=0}^{k_g-1} (\bar{h}_i^{(g)} - x_n^{(g)})^2 + (\bar{h}_j^{(g)} - \lambda_n^{(g)} x_n^{(g)} - p_n^{(g)})^2 \right),$$

$$\text{где } p_n^{(w)} = \frac{\bar{h}_i^{(w)} h_m^{(w)} - \bar{h}_j^{(w)} h_n^{(w)}}{h_{jn}^{(w)} - \bar{h}_j^{(w)}}; \quad \lambda_n^{(w)} = \frac{h_{in}^{(w)} - \bar{h}_i^{(w)}}{h_{jn}^{(w)} - \bar{h}_j^{(w)}},$$

$$x_n^{(w)} = p_n^{(w)} \frac{-2\lambda_n^{(w)} a_{2ij} - a_{3ij} \pm \sqrt{a_{3ij}^2 - 4a_{2ij}a_{1ij}}}{2(a_{1ij} + a_{2ij}(\lambda_n^{(w)})^2 + a_{3ij}\lambda_n^{(w)})}.$$

Дифференцируя последнее выражение по неизвестным $a_{1j} - a_{3j}$ получим систему нелинейных алгебраических уравнений относительно этих параметров, которая допускает лишь численное решение. Пусть мы нашли эти решения. Подстановка найденных значений $a_{1j} - a_{3j}$ в (1.62) и вычисление минимума функционала позволяют судить о величине вклада компонент i и j в классификацию фонем f и g .

В качестве критерия отбора пар компонент, которые следует

учитывать на стадии классификации фонем f и g , установим следующее условие:

$$\Delta_{ij}^{(fg)} < \frac{1}{C_F^2} \sum_{k=1}^N \sum_{l=1}^N \Delta_{kl}^{(fg)}. \quad (1.64)$$

На этом шаге можно остановить селекцию компонент, но возможно продолжение рассуждений с целью построить более глубокие функциональные зависимости и минимизировать количество компонент участвующих при различении классов. Для последовательного изложения дальнейших операций, процедуру отбора компонент (1.62)–(1.64) полинома (1.61) назовем первым шагом селекции компонент.

Допустим, что на первом шаге селекции пар компонент для фонем f и g удовлетворяют условию (1.64). Для значений полиномов, образованных этими парами сегментов, введем обозначения

$$x_1^{(w)}(2) = \varphi(x_{i_1}^{(w)}, x_{j_1}^{(w)}), \dots, x_{D_d}^{(w)}(2) = \varphi(x_{i_D}^{(w)}, x_{j_D}^{(w)}),$$

где для упрощения записи опущен индекс номера наблюдения.

Новые переменные $x_1^{(w)}(2), x_2^{(w)}(2), \dots, x_{D_d}^{(w)}(2)$ позволяют провести второй шаг селекции. Для этого необходимо подставить их в полином вида (1.61) вместо переменных $x_{ni}^{(w)}, x_{nj}^{(w)}$, затем провести операцию дифференцирования нового функционала (1.62) и численного решения полученной системы алгебраических нелинейных уравнений для вычисления неизвестных коэффициентов $a_{1j}(2) - a_{3j}(2)$ и найти значения минимума $\Delta_{ij}^{(fg)}(2)$ на втором шаге селекции. Сегменты, которые удовлетворяют условию вида (1.64), образуют входные данные для следующего шага селекции. Продолжая описанную методику, можно продолжать селекцию далее.

Критерием остановки селекции на $(n-1)$ шаге является условие

$$\frac{1}{D_{n-1}^{(fg)}} \sum_{k=1}^N \sum_{l=1}^N \Delta_{kl}^{(fg)}(n-1) > \frac{1}{D_n^{(fg)}} \sum_{k=1}^N \sum_{l=1}^N \Delta_{kl}^{(fg)}(n),$$

где n — номер шага селекции.

Пусть селекция была прекращена на n -ом шаге, тогда на этом шаге селекции были определены полиномы

$$x_i^{(w)}(n) = \phi(x_{i_1}^{(w)}(n-1), x_{i_2}^{(w)}(n-1), \dots, x_{i_D}^{(w)}(n)) = \phi(x_{i_1}^{(w)}(n-1), x_{i_2}^{(w)}(n-1)).$$

Таким образом, в заключение вычислений получим $D_n^{(fg)}$ функциональных зависимостей между исходными спектральными оценками фонем.

Глава 2. ИДЕНТИФИКАЦИЯ ДИКТОРОВ

Прежде чем приступить к изложению моделей идентификации дикторов, введем классификацию задач, выдвигаемых перед этими моделями и типами параметров (признаков), на основе которых предполагается решить эти задачи.

Рассмотрим типы входных речевых данных. В первом случае от пользователя требуется произнести строго заданное ключевое слово или фразу (пароль). В зависимости от того, кем его признает система, ему предоставляется определенный набор услуг. Такой способ идентификации характерен для систем контроля доступа в помещения, вычислительные комплексы и т.д. Во втором случае ключевое слово неизвестно пользователю до момента подхода к идентификатору. По требованию пользователя, идентификатор случайным образом выбирает слово или фразу из некоторого ограниченного словаря, которую пользователь должен произнести. Этот способ также характерен для систем контроля доступа, но с более жесткими условиями доступа, чем в первом случае. И в третьем случае требуется проводить идентификацию по абсолютно неизвестной фразе. Такая ситуация встречается в криминалистике.

Рассмотрим типы параметров используемых для идентификации. Как показывает практика, проявление индивидуальности голоса человека следует искать в двух основных группах признаков. Первая группа связана с анатомическими особенностями вокального тракта, а вторая — с уникальным механизмом приведения его в действие (артикуляционной деятельностью), который обусловлен работой центральной нервной системы.

Анатомические особенности человека проявляются в известной модели речевого тракта [14], включающей в себя передаточную функцию резонансной системы и генератор импульсов сигнала возбуждения. Передаточная функция характеризует индивидуальную геометрическую форму полостей речевого аппарата. Основными параметрами здесь выступают характеристики формантных областей: средняя частота, частотный диапазон,

энергия. Частота импульсов возбуждения находится в прямой зависимости от колебаний голосовых складок, которые, в свою очередь, зависят от длины, толщины и натяжения последних. Основным параметром здесь является средняя частота основного тона.

Если первая группа признаков отражает статические свойства речеобразующего тракта, то вторая группа должна описать его поведение во времени. Согласно существующему предположению, исходным и основным этапом в организации процесса речеобразования является управляемый центральной нервной системой человека комплекс программ артикуляционных движений, соответствующий тому сообщению, передача которого планируется в данный момент времени [6]. Комплекс программ артикуляционных движений является индивидуальным и определен предшествующим опытом человека. Здесь под артикуляционной подразумевается программа, которая содержит правила произнесения определенных структур. Эти правила относятся к управлению интонацией речи, ее ритмикой, ударением, громкостью, т.е. к управлению просодическими характеристиками речи.

Для расчета параметров, описывающих артикуляционную динамику речи, могут быть использованы методы спектрально-временного анализа данных. Однако необходимо отметить такую особенность расчета просодических параметров, как их жесткая связь с лексическим и синтаксическим контекстом исследуемой фразы. Это требует комплексного применения как средств лингвистического анализа, так и параметрических методов обработки, что явно определяет сложность анализа данных характеристик. При этом основной задачей является установление прямой связи между деятельностью речеобразующего аппарата (динамикой его артикуляционных движений) и характеристиками спектральной картины потока речи.

Итак, системы идентификации могут использовать статические и динамические признаки и их смеси.

Еще один практически важный аспект создания систем идентификации дикторов состоит в том, что множество дикторов, на котором решается задача, может быть открытым или замк-

нутым. На замкнутом множестве, от системы, по исходной речевой посылке, требуется принять решение о том, кто из дикторов этого множества произнес фразу. На открытом множестве допускается решение, что ни один из известных системе дикторов не произносил данную фразу.

В заключение поместим классификационные признаки систем идентификации дикторов в сводную таблицу 2.1.

Таблица 2.1

Признаки классификации систем идентификации дикторов

Тип фразы	Тип множества дикторов	Тип признаков идентификации
Ключевая фраза известная и пользователю и системе	Открытое	Статические
Ключевая фраза известная системе, но неизвестная пользователю	Замкнутое	Динамические
Фраза известная пользователю, но неизвестная системе	—	Смесь статических и динамических признаков

До настоящего времени доминирующую позицию в области идентификации по известной ключевой фразе занимают системы, которые основаны на моделях, использующих статические свойства речи. Такие системы начали развиваться в семидесятых годах и связаны с работами фирмы Texas Instruments [42, 43]. В них в качестве признаков были использованы мощности шестнадцати узкополосных фильтров равномерно расположенных в диапазоне от 300 до 3 000 Гц при произношении парольной фразы. В качестве эталона используется нормальное распределение амплитуд фильтров при допущении взаимной независимости этих амплитуд. Более подробное описание этой системы можно найти в работе [44]. Отметим, что в этом подходе слово используется как целое, т.е. не проводится сегментация парольного слова на более мелкие интервалы, которые могут нести информацию об индивидуальности голоса.

Логическим продолжением предыдущего подхода является другой распространенный метод идентификации, основанный на векторном квантовании и кепстральном приближении [45, 46],

который использует признаки последовательности парольных слов, взятых как целые. Здесь логарифмический спектр аппроксимируется в виде

$$\log|S(w)| \cong \sum_{k=-p}^p c(k) e^{-jkw}, \quad (2.1)$$

где $c(k)$ — кепстральные коэффициенты.

Дистанция между двумя кепстрами задана формулой

$$d(c_i, c_r) = \sum_{k=1}^n (1+k)^2 (c_i(k) - c_r(k))^2, \quad (2.2)$$

а дистанция между двумя последовательностями векторов кепстральных коэффициентов $\{c_i\}$ и $\{c_r\}$ определяется как среднее дистанций (2.2)

$$D(\{c_i\}, \{c_r\}) = \frac{1}{N} \sum_{k=1}^N d(c_{ik}, c_{rk}), \quad (2.3)$$

где одна из последовательностей $\{c\}$ определяет обучающую последовательность, а другая — распознаваемую последовательность; N — длина последовательности. Для идентификации диктора из всего множества эталонов выбирается такой эталон, который соответствует максимуму (2.3).

Экспериментальные результаты, полученные для этих систем, показывают, что системы обеспечивают в условиях небольших (10–20 дикторов) замкнутых множеств дикторов и при соотношении сигнал/шум не менее 20 дБ точность около 87 % и 89 %, соответственно порядку изложения. Однако рост количества дикторов, на которых обучена система, или переход от замкнутых множеств дикторов к открытым множествам приводит к катастрофическому падению точности идентификации.

Следующий шаг развития моделей идентификации дикторов представляется очевидным. Он связан с разбиением высказывания на сегменты и классификации каждого сегмента относительно содержащихся в памяти эталонов сегментов для каждого диктора.

Можно выделить три типа такой сегментации:

1) разбиение высказывания на *a priori* заданное или произвольное количество сегментов равной длительности;

2) разбиение высказывания на сегменты в моменты времени изменения какого-либо признака или набора признаков [47];

3) разбиение высказывания на сегменты соответствующие фонемам [47, 48]. Последний способ сегментации требует участия системы автоматического распознавания фонем в процессе идентификации [49].

В дальнейшем основное внимание будет уделено перечисленным способам сегментации высказывания для задач текстонезависимой идентификации.

2.1. Алгоритмы текстонезависимой системы идентификации диктора (ТСИД) для произвольных типов сегментации

2.1.1. Векторное квантование

Одним из эффективных способов идентификации диктора является векторное квантование (ВК). Суть этого способа заключается в том, что параметрическое пространство характеристик голоса диктора разбивается на конечное количество ячеек (кластеров). В кластер выделяется область пространства, где плотность попадания векторных характеристик голоса (векторов признаков) более высока, чем в окружающей области. Такое разбиение параметрического пространства является индивидуальным для каждого диктора.

Идентификация говорящего может быть построена двумя способами: либо по речевому сообщению строятся кластеры, и сравнивается их распределение в параметрическом пространстве с распределением, полученным при обучении (эталонным распределением); либо мы наблюдаем за процессом смены кластеров и принимаем решение: насколько такой процесс может принадлежать эталонному процессу, порождаемому голосом диктора.

При идентификации диктора можно хранить все вектора признаков, полученные при обучении, и в дальнейшем сравнивать поступающие вектора признаков с хранящимися в па-

мяти. В этом случае каждый обучающий вектор является кластером. На практике такой подход не применим, так как количество векторов в обучающей выборке может оказаться большим. Следовательно, необходимо найти метод, позволяющий сократить количество хранимых векторов без существенной потери точности идентификации.

Предположим, что $\mathbf{x} = [x_1, x_2, \dots, x_n]$ — n -мерный вектор признаков. Операция ВК вектору \mathbf{x} ставит в соответствие некоторый другой вектор y_i . Определим данную операцию как $y = Q(x)$, где $Q(\bullet)$ — это оператор квантования. Вектор y_i , называемый кодовый вектор, выбирается из конечного множества векторов $Y = \{y_1, y_2, \dots, y_S\}$. Множество Y называется кодовой книгой, а количество векторов в кодовой книге S называется ее размером. Для создания кодовой книги n -мерное пространство векторов признаков делят на S частей, каждая из которых описывается своим кодовым вектором y_i . На рисунке 2.1 показан пример разбиения двухмерного пространства пятью кодовыми векторами.

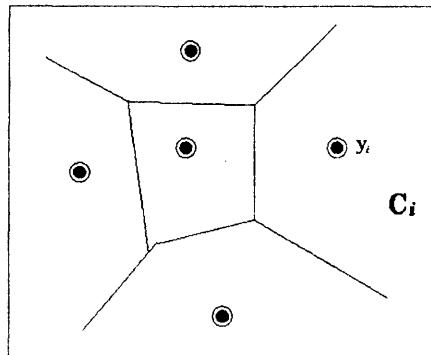


Рис. 2.1. Пример кодовой книги в двухмерном пространстве

Когда x квантуется как y , это приводит к ошибке квантования $d(x, y)$. В зависимости от приложений в качестве $d(x, y)$ выбирают различные функции, в зарубежной литературе называемые «distortion measure». Наиболее популярными функциями оценки ошибки квантования являются: квадрат расстояния

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2, \text{ где } n \text{ — размерность вектора; расстояние}$$

Махalanобиса $d(x, y) = (x - y)' W (x - y)$, где W — положительно определенная матрица .

Процедура вычисления подходящей кодовой книги Y при данной обучающей выборке X состоит в минимизации функционала

$$D = \frac{1}{N} \sum_{i=1}^N d(x_i, Q(x_i)),$$

который определяет среднюю ошибку квантования, здесь N — количество обучающих векторов.

Поскольку количество векторов в обучающей выборке на практике гораздо больше размера кодовой книги и зависимость средней ошибки D от кодового вектора y_i — нелинейная, то прямой процедуры вычисления кодовых векторов по обучающей выборке не существует. На данный момент предложено большое число итерационных процедур позволяющих строить множество Y . Одними из наиболее известных являются алгоритмы обучения на основе k -средних (k -means) и ЛБГ (LBG), названный так по первым буквам фамилий ее создателей.

2.1.1.1. Решающие правила

Рассмотрим решающие правила, использующиеся обычно при решении задачи идентификации дикторов. Если необходимо произвести классификацию одного вектора, можно использовать правило ближайшего кодового вектора (минимального расстояния). При идентификации дикторов задача усложняется тем, что вместо одного вектора имеется N векторов. Для последовательности векторов может быть предложено больше критериев, чем для одного вектора. Наиболее часто встречающимися на практике являются два критерия: по минимальному среднему расстоянию и на основе мажоритарного решения по каждому вектору. Приведем их определение в случае решения задачи идентификации на замкнутом множестве дикторов

Пусть для L дикторов сформированы кодовые книги Y_k и $\{x_i\}_1^N$ — входная последовательность векторов, тогда при использовании первого критерия вычисляются средние расстояния от входной последовательности до каждого объекта:

$$s_k(x_i) = \min_{y_j \in Y_k} [d(x_i, y_j)];$$

$$S_k = \frac{1}{N} \sum_{k=1}^N s_k(x_i). \quad (2.4)$$

В качестве диктора, которому принадлежит данная последовательность, выбирается тот, расстояние S_k для которого минимально:

$$\hat{g}(\{x_i\}_1^N) = \operatorname{argmin}_{1 \leq k \leq L} S_k.$$

Второй критерий основан на мажоритарном решении по каждому вектору:

$$m_i^k = \begin{cases} 1, & \text{если } s_k(x_i) \text{ минимально } \forall k = 1, 2, \dots, L \\ 0, & \text{в другом случае} \end{cases},$$

$$M_k = \sum_{i=1}^N m_i^k.$$

M_k представляет собой количество векторов, отнесенных к кодовой книге объекта k . В качестве объекта (диктора), которому принадлежит данная последовательность, выбирается тот, к чей кодовой книге отнесено наибольшее число векторов:

$$\hat{g}(\{x_i\}_1^N) = \operatorname{argmax}_{1 \leq k \leq L} M_k.$$

2.1.1.2. Алгоритм обучения на основе k -средних, ЛБГ алгоритм

Даны N векторов обучающей выборки $\{x_1, x_2, \dots, x_N\}$, необходимо разделить пространство X на S непересекающихся областей. Каждая из областей описывается своим центральным вектором y_i . В этом случае алгоритм вычисления кодовой книги Y на основе k -средних выглядит следующим образом:

1) случайным образом выбираются S векторов из обучающей выборки, которые и составляют начальную кодовую книгу;

2) в каждой области вычисляем новое значение центра классера

$$y_i \leftarrow \frac{1}{N_i} \sum_{x_j: Q(x_j) = y_i} x_j, \quad 1 \leq i \leq S;$$

3) шаг 2 повторяется до тех пор, пока не будет выполнен некоторый критерий окончания итераций.

В качестве функции квантования, как правило, выбирается критерий минимального расстояния:

$$Q(x_i) = y \cdot \operatorname{argmin}_j (d(x_i, y_j)). \quad (2.5)$$

Для данного алгоритма доказана его сходимость к минимуму функции D . Также исследована скорость сходимости, показано, что скорость сходимости превышает данную величину для метода градиентного спуска.

Несмотря на то, что процедура обучения на основе k -средних работает достаточно хорошо, Linde, Buzo и Gray предложили более эффективный алгоритм, названный ЛБГ [50]. Этот алгоритм стартует с кодовой книги, состоящей из одного вектора, являющегося средним вектором по всей обучающей выборке. Затем размер кодовой книги удваивается путем замены каждого вектора на два других: $y+p$ и $y-p$, путем добавления и вычитания небольшого вектора ошибки p . Полученная кодовая книга оптимизируется с использованием стандартного алгоритма на основе k -средних. Данные два шага повторяются до тех пор, пока средняя величина ошибки квантования не станет меньше некоторого порога, либо не будет выполнен какой-либо другой критерий окончания итераций.

Вышеприведенные рассуждения приводят к следующему алгоритму:

1) создаем начальную кодовую книгу $Y = \{y_1\}$,

$$\text{где } y_1 = \frac{1}{N} \sum_{i=1}^N x_i, \quad k = 1;$$

2) удваиваем размер кодовой книги, формируя на основе одного вектора два:

$$\begin{aligned} y_i &\leftarrow y_{i-k} - p, \quad k < i \leq 2k, \\ y_i &\leftarrow y_{i-k} - p, \quad 1 < i \leq k, \\ k &\leftarrow 2k, \end{aligned}$$

где p — небольшой вектор ошибки, выбираемый пропорциональным величине средней ошибки квантования;

3) вычисляем величину средней ошибки квантования

$$D_m = \frac{1}{N} \sum_{i=1}^N d(x_i, Q(x_i)),$$

где m — номер шага;

4) если $D_{m-1} - D_m < E$, то переходим к шагу 5, иначе в каждой области вычисляем новое значение центра кластера

$$y_i \leftarrow \frac{1}{N_{i-x_i \in O_i}} \sum_{x_j \in O_i} x_j,$$

и затем переходим к шагу 3;

5) если размер кодовой книги $k < S$, то переходим к шагу 2, иначе алгоритм считается завершенным.

В качестве функции квантования также выбирается минимальное расстояние (2.5).

Как видно из алгоритма, размер кодовой книги после завершения алгоритма всегда является степенью числа 2. Если это нежелательно, то в пункте 2 некоторые векторы не надо разделять. Кроме критерия достижения заданного количества векторов в пункте 5 возможно применение других критериев, например, достижения заданной величины средней ошибки квантования.

Проведенные в [50] практические исследования по нахождению зависимости вероятности верной идентификации для кодовых книг разного размера отображены в таблице 2.2.

Таблица 2.2

Результаты идентификации для кодовых книг различного размера

Порядок модели	ЛБГ
4	49,5
8	56,0
16	64,1
32	70,2

2.1.1.3. Метод обучения векторного квантования (ОВК)

Кодовая книга, построенная с помощью алгоритма ЛБГ, является оптимальной в том смысле, что средняя ошибка квантования минимальна. Она является оптимальной для использования, например, в кодировании речевого сигнала. Но в идентификации дикторов решение о классификации зависит не только от ошибок квантования для векторов из того же класса, но и от ошибок квантования векторов, принадлежащих другим классам. Поскольку в алгоритме ЛБГ используются только вектора из одного и того же класса, то построенная им кодовая книга является не самой оптимальной для использования в задачах идентификации дикторов. Идеальная кодовая книга должна строиться исходя из критерия минимизации ошибки классификации векторов из признакового пространства. Кохонен [51] предложил несколько алгоритмов модификации кодовых книг, позволяющих уменьшить величину ошибки классификации. Далее эти алгоритмы будут описаны как ОВК1, ОВК2 и ОВК3. Все три алгоритма используют уже созданные кодовые книги, которые могут быть получены, например, с помощью алгоритма ЛБГ.

Первый из описываемых алгоритмов назовем ОВК1. Основная идея данного метода заключается в том, что если вектор из обучающей выборки классифицирован верно, то ближайший к нему вектор из всех кодовых книг сдвигается в сторону данного обучающего вектора (становится к нему ближе), в противном случае кодовый вектор сдвигается в противоположную сторону (становится дальше).

Предположим, что имеется L наборов обучающих выборок $\{X_k\}_{k=1}^L$, $X_k = \{x_{kj}\}_{j=1}^{N_k}$ соответствующих разным объектам (дикторам). Предположим, также, что имеются L кодовых книг $\{Y_k\}_{k=1}^L$, $Y_k = \{y_{kl}\}_{l=1}^{N_k}$, сформированных алгоритмом ЛБГ по имеющимся выборкам X_k . Используя эти кодовые книги, можно классифицировать каждый вектор из обучающих выборок и найти

ближайший к нему вектор y_{ki} какой-либо кодовой книги Y_k . В этом случае можно уменьшить количество ошибок идентификации путем модификации (доводки) кодовых книг по следующему алгоритму:

- случайным образом выбирается вектор x_{ij} из какой-либо обучающей выборки X_i , и по всем кодовым книгам $\{Y_k\}_{k=1}^L$ находится ближайший кодовый вектор y_{kl} , принадлежащий кодовой книге Y_k . Если $k = i$, т.е. объекты, соответствующие обучающему и кодовому векторам, совпадают, то кодовый вектор модифицируется следующим образом:

$$y_{kl} \leftarrow y_{kl} + a(x_{ij} - y_{kl}),$$

в противном случае кодовый вектор модифицируется следующим образом:

$$y_{kl} \leftarrow y_{kl} - a(x_{ij} - y_{kl}),$$

где a — это достаточно малый параметр, задающий скорость обучения. При практических реализациях он выбирается равным 0,01 и в последующем монотонно убывает до нуля.

В случаях, когда обучающие выборки различных объектов пересекаются, невозможно точно разделить объекты. Следовательно, задачей является построение разделяющих границ (решающих правил), минимизирующих количество ошибочно классифицированных векторов из обучающих выборок. Этот результат может быть достигнут размещением разделяющих границ в точках, где плотности вероятностей, соответствующие различным объектам, пересекаются. Выбор параметров разделяющих правил по данному правилу является оптимальной байесовской оценкой параметров разделяющих функций. Кохонен доказал, что вышеупомянутый алгоритм обучения приводит к кодовым векторам, описывающим разделяющие правила, которые аппроксимируют разделяющие границы, оптимальные в смысле байесовской оценки параметров.

Предыдущий алгоритм может быть достаточно легко модифицирован для улучшения соответствия байесовской оценки параметров, исходя из следующих рассуждений. Для обучающего вектора x_{ij} находится два ближайших к нему кодовых вектора

из всех кодовых книг y_{kl} и y_{qm} . Если один и только один из векторов y_{kl} , y_{qm} принадлежит кодовой книге i , т.е. $k = i$ или $q = i$ (для определенности предположим что $q = i$), и если вектор x_{ij} лежит в некоторой, достаточно малой, окрестности средней точки лежащей между векторами y_{kl} и y_{qm} , то вектор y_{kl} отодвигается от вектора x_{ij} а вектор y_{qm} — придвигается к нему. Другими словами:

$$y_{kl} \leftarrow y_{kl} - a(x_{ij} - y_{kl}),$$

$$y_{qm} \leftarrow y_{qm} + a(x_{ij} - y_{qm}),$$

где a — это малый параметр, отвечающий за скорость обучения, монотонно стремящийся к нулю. Оптимальная величина окрестности средней точки зависит от объема обучающей выборки и в каждом случае определяется экспериментально. При небольшом объеме в качестве размера области обычно принимают число между $0,1d(y_{qm}, y_{kl})$ и $0,2d(y_{qm}, y_{kl})$. С увеличением объема обучающей выборки размер области выбирают меньшим, но, как правило, зависящим от расстояния между векторами.

Интуитивно понятно, что алгоритм ОВК2 использует локальные особенности для сдвига разделяющих границ к оптимальным байесовским. С другой стороны, данный алгоритм совершенно не учитывает глобальных изменений, привнесенных его действиями. На практике это выражается в том, что количество ошибок классификации при применении данного алгоритма поначалу падает, пока разделяющие границы смещаются в сторону оптимальных байесовских, но потом начинают увеличиваться. Таким образом, данный алгоритм следует применять лишь ограниченное количество итераций. Вводя дополнительную коррекцию путем комбинирования двух алгоритмов ОВК1 и ОВК2 Кохонен предложил алгоритм ОВК3, состоящий в следующем:

- для выбранного случайным образом вектора x_{ij} по всем кодовым книгам находятся два ближайших кодовых вектора y_{kl} и y_{qm} ; если все три вектора удовлетворяют двум условиям, опи-

санным в алгоритме ОВК2, то преобразования кодовых векторов выполняется по правилам, описанным в алгоритме ОВК2. В противном случае, если $k=1$ и $q=i$, то кодовые вектора преобразуются по следующему правилу:

$$y_{kl} \leftarrow y_{kl} + ea(x_{ij} - y_{kl}),$$

$$y_{qm} \leftarrow y_{qm} + ea(x_{ij} - y_{qm}),$$

где a — малый параметр, называемый скоростью обучения, так же как и в ОВК2; e — еще одна малая константа выбираемая из интервала от 0,1 до 0,5. На практике алгоритм ОВК3 является сходящимся, и формируемые им кодовые книги обеспечивают меньшую величину ошибки классификации векторов из обучающих выборок, чем кодовые книги формируемые алгоритмами ОВК1 и ОВК2.

2.1.2. Гауссовые смеси

В главе 1 (в связи с вопросами оценки распределения плотности вероятностей шума и речевого сигнала) мы упоминали гауссовые смеси (Gaussian Mixture Model — GMM), которые впервые были предложены для решения задачи оценки плотности вероятности случайной функции по ее выборке Розенблатом (1956) и Парзеном (1962). Практически сразу этот подход был использован для решения задач распознавания образов [52]. Однако первые попытки использования данного подхода в задачах обработки речи и идентификации дикторов относятся лишь к середине девяностых годов XX в. [53].

Данный подход основывается на моделировании плотности распределения вероятности появления некоторого вектора акустических параметров с помощью взвешенной суммы функций. В качестве таковых выбираются функции, локализованные в некоторой области пространства, и имеющие конечный интеграл при ее интегрировании по всему пространству. На практике чаще используются функции Гаусса.

Модель на основе гауссовых смесей используется для аппроксимации плотности вероятности распределения векторов \mathbf{x} из пространства наблюдений X . При построении модели на ос-

нове гауссовых смесей плотность вероятности $p(\mathbf{x})$ задается выражением:

$$p(\mathbf{x} | \Theta) = \sum_{i=1}^N \alpha_i p(\mathbf{x} | \mu_i, \Sigma_i), \quad (2.6)$$

где α_i — вес i -го гауссиана, сумма которых удовлетворяет условию

$$\sum_{i=1}^N \alpha_i = 1;$$

μ_i — вектор математического ожидания i -го гауссиана;

Σ_i — ковариационная матрица i -го гауссиана;

Θ — набор параметров гауссовой смеси, $\Theta = \{\alpha_i, \mu_i, \Sigma_i\}$;

$$p(\mathbf{x} | \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} \det \Sigma^{1/2}} \exp \{-(\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i)\}. \quad (2.7)$$

Выражение (2.7) содержит операции матричного умножения и обращения ковариационной матрицы. Данные операции достаточно сложны в вычислительном плане. Поэтому часто выдвигают предположения относительно независимости компонент вектора \mathbf{x} , которые позволяют привести ковариационную матрицу к диагональному виду. (Подобное предположение лишь незначительно снижает точность распознавания).

2.1.2.1. Решающее правило

Предположим теперь, что имеется группа из L дикторов, и для каждого из них сформирован набор параметров модели Θ_i , $i=1, \dots, L$. Задача состоит в определении диктора, чья модель имеет максимальную вероятность для наблюдаемой последовательности состояний $X = \{x_i\}_{i=1}^T$.

При решении задачи классификации одного вектора \mathbf{x} , используя правило Байеса, получим вероятность того, что этот вектор принадлежит диктору с номером i :

$$Pr(i | \mathbf{x}) = \frac{p(\mathbf{x} | \Theta_i)}{p(\mathbf{x})} Pr(i),$$

где $Pr(i)$ — априорная вероятность того, что неизвестный дик-

тор является диктором с номером i ; $p(x)$ — вероятность появления вектора из пространства наблюдений x .

Основываясь на принципе максимального правдоподобия, можно утверждать, что диктор, для которого данная вероятность будет максимальной и будет диктором, породившим вектор признаков x :

$$p(x|\Theta_i)Pr(i) > p(x|\Theta_j)Pr(j), \quad j=1, 2, \dots, L, \quad j \neq i.$$

Полагая, что в качестве неизвестного диктора любой из дикторов может выступать с равной вероятностью ($Pr(i)=1/L$), получаем решающее правило для одного вектора признаков:

$$\hat{g}_1(x) = \operatorname{argmax}_{1 \leq i \leq L} \{p(x|i)\}.$$

Переходя к рассмотрению последовательности наблюдений X и полагая, что вектора в этой последовательности не зависят один от другого, получаем функцию правдоподобия того, что данная последовательность произнесена диктором i :

$$P(X|\Theta_i) = \prod_{t=1}^T p(x_t|\Theta_i). \quad (2.8)$$

Повторяя вышеприведенные рассуждения, приходят к следующему решающему правилу:

$$\hat{g}(X) = \operatorname{argmax}_{1 \leq k \leq L} \{P(X|\Theta_k)\}.$$

На практике вместо функции $P(X|\Theta)$ используют логарифм этого выражения, что приводит к следующей форме решающего правила:

$$\hat{g}(X) = \operatorname{argmax}_{1 \leq k \leq L} \{\log(P(X|Y_k))\} = \operatorname{argmax}_{1 \leq k \leq L} \left\{ \sum_{t=1}^T \log(p(x_t|Y_k)) \right\}.$$

Процедура вычисления параметров гауссовых смесей на основе EM-алгоритма подробно рассмотрена в приложении 3.

2.1.3. Методы нормализации характеристик

Мы рассмотрели два метода текстонезависимой идентификации на замкнутом множестве дикторов: метод векторного квантования и метод, основанный на гауссовых смесях. При переходе

к решению задачи идентификации на открытом множестве дикторов на первый взгляд достаточно ввести порог на вычисляемую меру, используемую для принятия решения. Например, для методов, основанных на векторном квантовании, достаточно ввести порог на среднее расстояние от кодовой книги до анализируемой речи, при превышении которого принимать решение о том, что диктор системе неизвестен. На практике же данный подход работает плохо, приводит (как и в случае систем идентификации по парольной фразе) к катастрофическому паданию точности. Для решения этой проблемы были предложены методы нормализации вычисляемых характеристик.

Предполагается, что основной вклад в вариации голоса диктора вносят изменения его эмоционального состояния, состояния здоровья и типы каналов передающих речь. Для построения системы, устойчивой к этим изменениям, используют разные методы принятия решения на основе числовых характеристик, формируемых моделью диктора. Одним из наиболее распространенных и наиболее эффективных методов является метод нормализации числовых характеристик.

Первая из рассматриваемых норм получила название Z -норма. Данная норма состоит в приведении распределения средних расстояний к нормальному распределению с нулевым средним и единичной дисперсией по следующей формуле:

$$R_Z(S) = \frac{R(S) - \mu_s}{\sigma_s},$$

где $R_Z(S)$ это Z -нормированное расстояние от тестового произношения до модели диктора S ; $R(S)$ — расстояние от тестового произношения до модели диктора S ; μ_s — среднее значение расстояния $R(S)$, вычисленное на обучающей последовательности; σ_s — дисперсия расстояния $R(S)$, вычисленная на обучающей последовательности.

Данное преобразование не приводит к какому-либо повышению точности верификации по сравнению с методом индивидуального порога для каждой модели диктора. Единственное его преимущество состоит в возможности использования одного порога и возможности его адаптивно менять в зависимости от текущих потребностей.

Следующий метод получил название Н-норма (от английской

кого слова Handset). Данная норма является дальнейшим развитием Z -нормы на случай нескольких типов каналов:

$$R_z(S) = \frac{R(S) - \mu_{S,H}}{\sigma_{S,H}}.$$

В этом случае среднее значение и дисперсия зависят не только от диктора, но и от канала. Количество каналов обычно не превышает нескольких единиц. Например, при обработке речи, источником которой является телефонный канал, используют два типа каналов, соответствующих угольному и электретному микрофонам. Следует отметить, что использование в формуле типа канала порождает проблему его автоматического определения.

В заключение опишем метод, получивший название UBM. Данный метод состоит в введении еще одной модели диктора (U), для обучения которой используется речь не одного диктора, а большого числа различных дикторов разного пола, записанная в различных условиях. Нормированное расстояние находится по следующей формуле:

$$R_{UBM}(S) = R(S) - R(U),$$

где $R_{UBM}(S)$ — UBM-нормированное расстояние от тестового произношения до модели диктора S ; $R(S)$ — расстояние от тестового произношения до модели диктора S ; $R(U)$ — расстояние от тестового произношения до общей модели U .

Очевидно, что полученное расстояние уже не будет строго положительной величиной, тем не менее, то свойство, что чем меньше величина $R_{UBM}(S)$, тем больше уверенность, что тестовая последовательность произнесена диктором S — сохраняется. Дальнейшее развитие данного метода состоит в создании различных общих моделей для различных типов каналов и пола дикторов.

2.2. Алгоритмы ТСИД с распознаванием гласных звуков

Идея структуры ТСИД изображена на рисунке 2.2. Система должна обладать тремя видами деятельности: а) формирование

фонетических классов или фонетическое обучение; б) формирование распределений параметров голоса определенного диктора или обучение на диктора; в) идентификация.



Рис. 2.2. Структурная схема ТСИД, основанной на распознавании фонем

Блок фонетического обучения предназначен для сбора и обработки гласных фонем, произнесенных произвольными дикторами, а так же формирования представления признаков фонем, которое инвариантно относительно говорящего, т.е. по сути дела этот блок создает в пространстве параметров области, соответствующие гласным фонемам (области «а», «о», «и» на рис. 2.3).

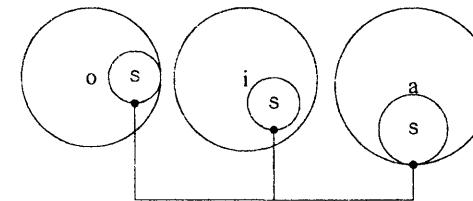


Рис. 2.3. Схематическое изображение областей, соответствующих фонемам в некотором параметрическом пространстве, включающих в себя подобласти, соответствующие фонеме диктора S

Блок обучения на диктора в области признаков каждой фонемы формирует подобласть, соответствующую индивидуальному произношению этой фонемы диктором, и создает связи между этими областями (подобласти диктора « s » в области каждой фонемы на рис. 2.3).

Блок идентификации проводит классификацию гласных фонем в произнесенной фразе и определяет диктора, сказавшего данную фразу.

2.2.1. Оценка точности ТСИД с распознаванием гласных звуков

Поставим следующую задачу: оценить точность работы ТСИД, в основу деятельности которой положен анализ индивидуальности произношения звуков при условии, что существует система способная их выделить.

На основе известной таблицы значимостей звуков речи для идентификации дикторов, представленной в работе [6], можно сделать вывод, что для эффективного опознания диктора достаточно использовать акустические характеристики гласных фонем.

Идея структуры ТСИД состоит в том, что с помощью системы распознавания фонем происходит классификация входного вектора признаков с разделением на семь классов: «а», «о», «у», «и», «ы», «э», «не гласная фонема». По результатам классификации входной вектор признаков либо сравнивается с эталонными параметрами соответствующего класса конкретного диктора, либо не рассматривается вообще, если он попал в класс «не гласная фонема».

Для теоретической оценки точности ТСИД была создана база записей гласных звуков произнесенных различными дикторами. Двадцать пять мужчин и двадцать пять женщин произносили гласные звуки: («а», «о», «у», «и», «ы», «э») по четыре раза каждый звук. Средняя длительность одного произношения составляла 0,4 с. Возраст дикторов мужчин составляла от 22 до 56 лет, возраст дикторов женщин от 19 до 47 лет.

Входной сигнал оцифровывался с частотой 10 кГц. После чего он сегментировался на участки, длительностью 0,05 с с шагом 0,025 с. На каждом сегменте вычислялся спектр сигнала путем вычисления ДПФ. Полученный спектр сглаживался. Для сглаживания использовался фильтр Баттервортса 5-го порядка с частотой среза 360 Гц. Такая частота среза фильтра выбиралась,

исходя из известных данных [13] о минимальной дистанции между формантами гласных звуков. Известно [35], что рекурсивный фильтр Баттервортса имеет нелинейный фазовый сдвиг, для устранения которого используется метод двухсторонней фильтрации, т.е. метод, при котором данные обрабатываются линейным фильтром, а его выходные данные, взятые в обратном направлении, пропускаются через тот же фильтр. В результате такой обработки для каждого сегмента был получен сглаженный спектр $F_{ijk}^{(f)}(\omega)$, где i — номер сегмента, j — номер произношения, k — номер диктора, f — номер фонемы, ω — частота.

Для каждого диктора и каждой гармоники сглаженного спектра были найдены математическое ожидание и среднеквадратическое отклонение (СКО):

$$M_k^{(f)}(\omega) = MO(F_{ijk}^{(f)}(\omega)) = \frac{\sum_{i,j} F_{ijk}^{(f)}(\omega)}{N_k}, \quad (2.9)$$

$$\sigma_k^{(f)}(\omega) = \sigma(F_{ijk}^{(f)}(\omega)) = \sqrt{\frac{\sum_{i,j} (F_{ijk}^{(f)}(\omega))^2}{N_k} - M_k^2(\omega)}, \quad (2.10)$$

где N_k — суммарное количество сегментов по всем произношениям f -й фонемы k -м диктором. В дальнейшем под АЧХ голосового тракта диктора мы будем понимать функцию (2.9).

На рисунке 2.4 приведены наиболее характерные графики амплитудно-частотных характеристик для двух дикторов мужчин (M_06 , M_07 — их порядковые номера в базе данных). На рисунке 2.5 представлены наиболее типичные графики амплитудно-частотных характеристик для двух дикторов женщин (F_07 , F_08 — их порядковые номера в базе данных). Сплошной линией на графиках показано математическое ожидание, а пунктирной линией показано СКО.

Кроме функций (2.9) и (2.10), вычисленные для каждого диктора, аналогичным образом были найдены функции $M^{\phi}(\omega)$ и $\sigma^{\phi}(\omega)$, вычисленные для каждой фонемы. Для их вычисления были использованы все произношения, всех дикторов одного пола. Результаты представлены на рисунке 2.6

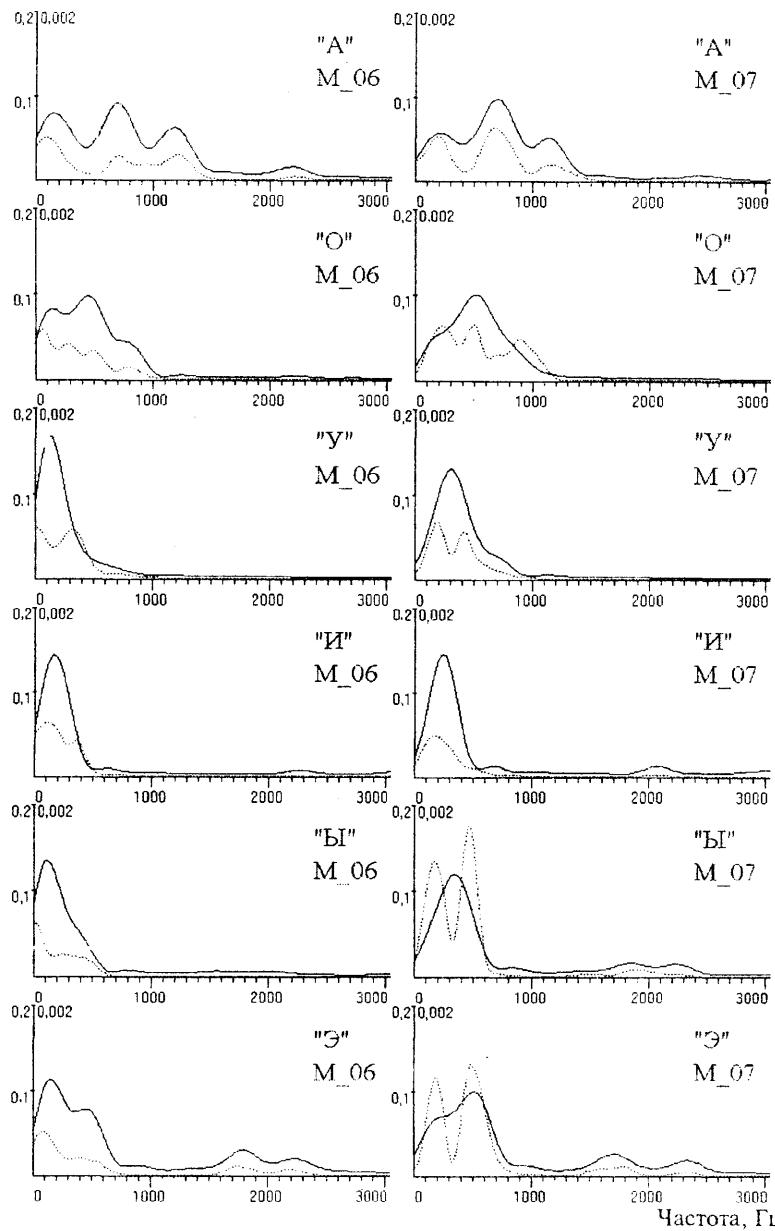


Рис. 2.4. Типичные АЧХ для двух дикторов мужчин

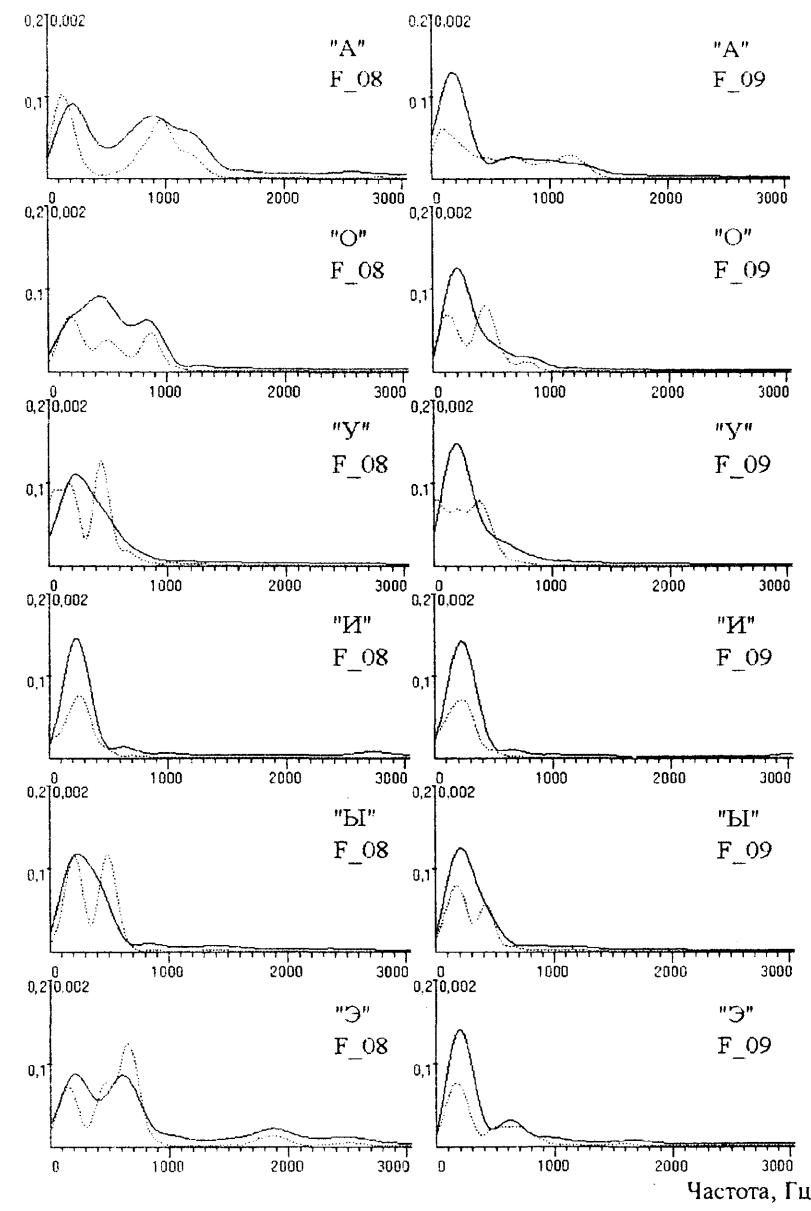


Рис. 2.5. Типичные АЧХ для двух дикторов женщин

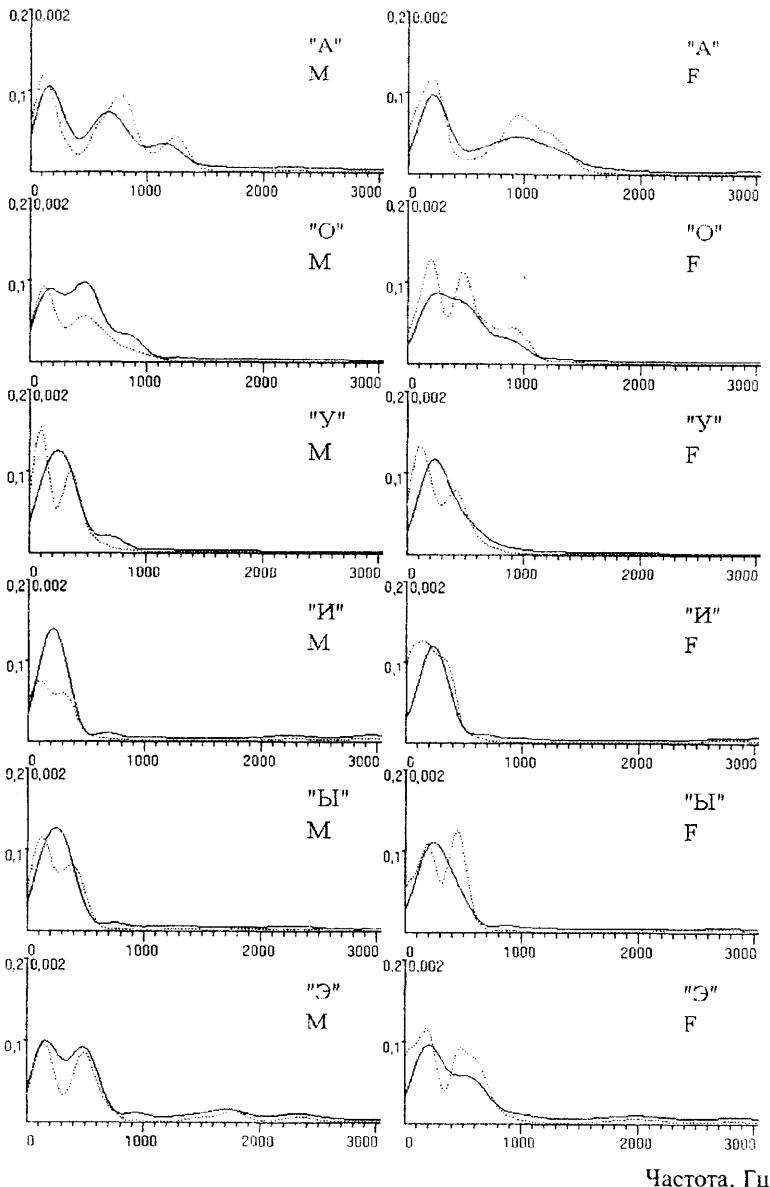


Рис. 2.6. Суммарное АЧХ: слева — для дикторов мужчин, справа — для дикторов женщин

Для определения формантных частот на основе полученных АЧХ использовался алгоритм приложения 3. В описываемых экспериментах количество гауссOID N выбиралось равным 8, а $\Delta\omega$ равным 100 Гц. Из 8 найденных частот ω , выбирались три значения, для которых значения A_i максимальны. После упорядочивания по частоте они и считались первой, второй и третьей формантами.

Для уточнения формантных частот на каждом интервале использовалось две операции. Во-первых, вычисляется функция (2.9) — $M_k(\omega)$ по всем произношениям данного диктора, и по ней «предсказываются» значения частот второй и третьей форманты — $\bar{\omega}_2, \bar{\omega}_3$. Во-вторых, на каждом сегменте вычисленные формантные частоты $\bar{\omega}_2$ и $\bar{\omega}_3$ сравниваются с «предсказанными». Если они удовлетворяют условию: $|\omega_i - \bar{\omega}_i| < \gamma$, где ω_i — предсказанное значение частоты i -ой форманты, то они добавляются в выборку, иначе — отбрасываются как ошибки. При получении практических результатов значение γ выбиралось равным 160 Гц.

По построенной выборке для частот и амплитуд формант находились их статистические характеристики: математическое ожидание и СКО. Некоторые типичные результаты приведены в таблицах 2.3 и 2.4, полный перечень представлен в работе [54].

Полученные данные позволяют найти оценку максимально-го количества классов, различаемых системой при заданной точности распознавания. Предположим, что значения каждого из параметров $\omega_2^{(f)}, \omega_3^{(f)}, a_2^{(f)}, a_3^{(f)}$ в t -ом измерении удовлетворяют неравенствам:

$$\begin{aligned} \bar{\omega}_2^{(f)} - 3\sigma(\omega_2^{(f)}) &\leq \omega_2^{(f)}(t) \leq \bar{\omega}_2^{(f)} + 3\sigma(\omega_2^{(f)}); \\ \bar{\omega}_3^{(f)} - 3\sigma(\omega_3^{(f)}) &\leq \omega_3^{(f)}(t) \leq \bar{\omega}_3^{(f)} + 3\sigma(\omega_3^{(f)}); \\ \bar{a}_2^{(f)} - 3\sigma(a_2^{(f)}) &\leq a_2^{(f)}(t) \leq \bar{a}_2^{(f)} + 3\sigma(a_2^{(f)}); \\ \bar{a}_3^{(f)} - 3\sigma(a_3^{(f)}) &\leq a_3^{(f)}(t) \leq \bar{a}_3^{(f)} + 3\sigma(a_3^{(f)}), \end{aligned}$$

Таблица 2.5

Интегральные характеристики фонем, вычисленные по всем дикторам.

Мужчины

Фонема	Форманта											
	1				2				3			
	Частота		Амплитуда		Частота		Амплитуда		Частота		Амплитуда	
	МО	СКО	МО	СКО	МО	СКО	МО	СКО	МО	СКО	МО	СКО
А	166,7	5,92	0,104	0,024	683,3	8,26	0,087	0,02	1155,1	9,5	0,043	0,011
Э	148,8	5,96	0,103	0,024	501,5	8,29	0,093	0,019	1000,8	8,76	0,014	0,003
И	214,5	5,43	0,144	0,021	691,4	7,14	0,011	0,004	1680,4	8,74	0,01	0,003
О	145,3	5,87	0,091	0,024	488,6	9,27	0,101	0,02	865,5	8,27	0,036	0,01
У	260,8	8,95	0,139	0,019	698,9	8,49	0,024	0,008	1050,4	7,24	0,019	0,002
Ы	253,3	7,81	0,138	0,019	740,9	9,12	0,013	0,004	1148,0	11,6	0,009	0,003

Таблица 2.6

Интегральные характеристики фонем, вычисленные по всем дикторам.

Женщины

Фонема	Форманта											
	1				2				3			
	Частота		Амплитуда		Частота		Амплитуда		Частота		Амплитуда	
	МО	СКО	МО	СКО	МО	СКО	МО	СКО	МО	СКО	МО	СКО
А	214,5	8,0	0,103	0,033	966,3	15,8	0,056	0,029	1344,6	15,2	0,035	0,021
Э	198,9	9,2	0,108	0,03	582,4	13,5	0,063	0,033	999,3	14	0,013	0,006
И	230,5	10,7	0,131	0,028	653,6	14,7	0,015	0,006	1016,9	11,6	0,008	0,002
О	216,9	9,7	0,103	0,03	482,7	13,5	0,093	0,032	898,6	14,4	0,034	0,024
У	243,2	12,0	0,126	0,027	574,3	12,9	0,042	0,018	939,5	15,6	0,01	0,004
Ы	252,8	13,5	0,122	0,028	543,9	15,6	0,044	0,037	892,5	14,1	0,011	0,004

Используя данные таблиц 2.5 и 2.6, оценим размеры формантных областей для каждой фонемы.

Зададим параметр — вероятность правильной классификации диктора по одной фонеме — E . Используя данные таблиц 2.3 и 2.4, можно оценить размеры формантных областей одного диктора для каждой фонемы с учетом параметра E . Разделив размер области фонемы на размер области той же фонемы для определенного диктора, получим количество дикторов, которое мы можем различить по одной фонеме. В таблице 2.7 приведены значения количества дикторов по всем фонемам для различного значения параметра E .

где $\bar{a}_2^{(f)}$, $\bar{a}_3^{(f)}$ — математические ожидания амплитуд, соответственно, второй и третьей формант, фонемы f .

Таблица 2.3

Характеристики фонем для одного диктора. Мужчины

Фонема	Форманта											
	1				2				3			
	Частота		Амплитуда		Частота		Амплитуда		Частота		Амплитуда	
	МО	СКО	МО	СКО	МО	СКО	МО	СКО	МО	СКО	МО	СКО
А	210,3	6,1	0,104	0,027	640,5	12	0,038	0,015	1048,1	13,3	0,048	0,019
Э	194,7	6,6	0,108	0,024	580,6	9,8	0,061	0,025	987,1	11,2	0,013	0,005
И	221,4	5,5	0,131	0,023	640,8	9,9	0,017	0,008	1760	10,5	0,01	0,003
О	302,5	9,1	0,103	0,024	663,1	13,0	0,061	0,023	947,3	11,6	0,021	0,01
У	240,4	8,7	0,128	0,021	585,3	12,0	0,043	0,021	959,7	11,6	0,011	0,005
Ы	246,3	9,3	0,124	0,024	662,8	12,4	0,032	0,017	1277,8	11,3	0,009	0,002

Результаты формантного анализа голосов дикторов мужчин и отдельно голосов дикторов женщин приведены в таблицах 2.5 и 2.6.

Таблица 2.7

Оценка количества идентифицируемых дикторов при заданной точности по одной фонеме

Фонема	Вероятность правильной классификации диктора				
	0,6	0,70	0,80	0,90	0,98
А	24	16	16	4	1
Э	8	2	2	2	1
И	18	18	12	4	2
О	16	16	1	1	1
У	6	2	2	2	1
Ы	8	4	2	1	1

Рассмотрим случай, когда для классификации используется последовательность фонем длиной L . Предположим, что правильная классификация происходит, если не менее половины гласных звуков будет классифицирована верно. Если предположить, что для всех фонем мы имеем одни и те же значения вероятностей: p — вероятность классификации диктора по одной фонеме, это значение задано параметром E , \bar{p} — вероятность ошибки, когда мы выбираем неверного диктора. Тогда выражение для вероятности правильной классификации V можно записать в виде

$$V = \sum_{i=1}^{\lfloor L/2 \rfloor} C_L^{\lfloor L/2 \rfloor - i} p^{\lfloor L/2 \rfloor + i} \bar{p}^{\lfloor L/2 \rfloor - i}, \quad (2.11)$$

где $\lfloor \cdot \rfloor$ — означает операцию взятия целой части числа; $C_L^{\lfloor L/2 \rfloor - i}$ — число сочетаний.

В таблице 2.8 приведены вычисленные значения вероятности распознавания для некоторых длин последовательностей.

Заметим, что так как в формуле (2.11) используется операция взятия целой части, полученная последовательность $\{V_i(L)\}$ для любого значения E не будет монотонной. Монотонными будут последовательности вида: $\{V_i(2L)\}$ и $\{V_i(2L - 1)\}$.

Таблица 2.8

Оценка зависимости точности идентификации от длины последовательности звуков

Длина последовательности звуков	Точность идентификации, %					
	L / E	60	70	80	90	98.9
1	0,6	0,7	0,8	0,9	0,989	
2	0,36	0,49	0,64	0,81	0,979	
3	0,64792	0,784	0,8958579	0,97182	0,999	
4	0,47514	0,652	0,8190484	0,94748	0,999	
6	0,54417	0,744	0,9007925	0,98374	1	
8	0,59384	0,806	0,9432139	0,9944	0,999	
10	0,63275	0,849	0,9665297	0,99763	0,999	

Необходимо отметить, что оценки являются несколько завышенными, поскольку не учитывают вероятность правильно распознавания фонемы, которая зависит от конкретной системы распознавания фонем, используемой для решения задачи идентификации.

Общий итог проведенных оценок следующий: чем более точно мы способны классифицировать гласный звук относительно множества дикторов, тем менее длинная последовательность этих звуков необходима для идентификации с заданной точностью.

2.2.2. Обучение на диктора

Рассмотрим работу блока обучения на диктора. Пусть в результате сегментации выделены интервалы речи, которые, с одной стороны, обладают основным тоном, а с другой стороны, их интенсивность больше, чем средняя интенсивность речи, взятая по фразе в целом. Эти признаки указывают на то, что выделенный интервал занят гласной фонемой. Окончательное решение принимается при сравнении формантного состава данного звука с формантными составами гласных, содержащихся в памяти системы, на основе метода сравнения формантных наборов, описанного в приложении 3.

Пусть на некотором сегменте $[T_0, T_1]$ найдена последовательность формантных представлений. Эта последовательность должна удовлетворять двум условиям:

- 1) любые соседние элементы последовательности должны быть схожи между собой, т.е. $\xi(L_t, L_{t+k}) > Q$ для любых t и k , удовлетворяющих условиям $T_0 \leq t \leq T_1$, $T_0 \leq t+k \leq T_1$, где $\xi(L_t, L_{t+k})$ — мера близости формантных наборов, рассчитанная в соответствии с (П.3.9); Q — некоторый порог (параметр модели);
- 2) каждый из элементов последовательности имеет максимальную степень схожести с одной и той же фонемой, содержащейся в памяти, т.е.

$$w = \underset{f}{\operatorname{argmax}} \xi(L_t, \theta_f)$$

является постоянной величиной для любого момента времени, удовлетворяющего условию $T_0 \leq t \leq T_1$, θ_f — эталон формантного набора фонемы f .

Если же среди всей последовательности найдутся такие соседние элементы, которые являются не сходными, то исходная последовательность разбивается на соответствующее число подпоследовательностей, и предполагается, что каждая из подпоследовательностей определяет фонетическое состояние. (Все они проходят проверку на условие 2).

На основе формантных наборов, для которых оба условия выполнены, формируется банк данных. Этот банк является исходным материалом для формирования эталонов фонем дикторов.

Необходимо отметить, что каждой фонеме диктора может соответствовать не один эталон. Этalonы могут быть сформированы из формантных наборов с равным количеством формант, а затем для каждой форманты определено среднее значение и дисперсия амплитуды и частоты. Этalonы могут быть сформированы с помощью метода векторного квантования, где в качестве расстояния участвует мера (Приложение 3).

2.2.3. Решающее правило и тестирование ТСИД

На основании данных о параметрах основного тона произнесенной фразы, происходит первый уровень сортировки голосов дикторов [55], т.е., если средняя частота ОТ в фразе равна $\omega_{\text{от}}$, то из базы данных выбираются те дикторы, частота ОТ которых удовлетворяет условию $|\omega_{\text{от}} - \omega_{\text{от},i}| < 40$ Гц, где $\omega_{\text{от},i}$ — средняя частота ОТ i -го диктора.

Пусть на протяжении фразы найдено K последовательностей, которые классифицированы как гласные фонемы и пусть на протяжении этих последовательностей найдены средние значения амплитуд и частот формант. Меру того, что полученные значения формант данной последовательности соответствуют какому-либо эталону i фонемы f диктора d получим исходя из формулы (П.3.9). Значение $m_i^{(d)} = \max_{f,u} \xi(L_t, \theta_{fu}^{(d)})$ определяет величину схожести характеристики последовательности наборов фонем L_t с одним из эталонов какой-либо фонемы диктора d , которая с максимальной вероятностью схожа с фонемой, представленной последовательностью. Величина схожести всей фразы с фразами диктора d определяется в виде среднего значения меры сходства последовательностей

$$\Xi^{(d)} = \frac{\sum_{t=1}^K k_t m_t^{(d)}}{\sum_{t=1}^K k_t}, \quad (2.12)$$

где k_t — длительность t -ой последовательности в единицах количества формантных наборов.

Максимальное значение меры определит индекс диктора, произносившего данную фразу.

Для проведения экспериментов над описанной ТСИД, где эталоны формировались по принципу равенства количества формант в исходных формантных наборах, была использована база, содержащая монологи 100 дикторов мужчин и женщин. Монологи были записаны в телефонном канале с частотой оциф-

ровки 8 кГц и формате A-Law и отношением сигнал/шум не менее 10 дБ.

В качестве управляемых величин использовались длительности тестового и обучающего сигналов. Из рисунка 2.7. видно, что подобные системы имеют практическую ценность при длительностях тестирующего и обучающего сигнала более 30 с.

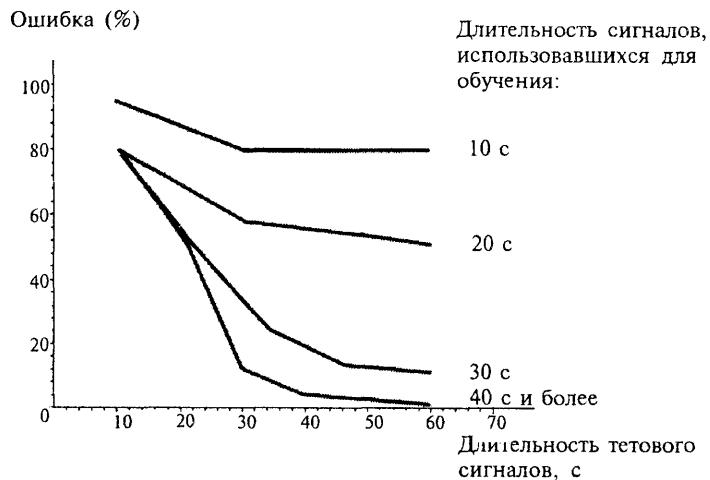


Рис. 2.7. Результаты экспериментальных исследований над ТСИД, основанной на распознавании гласных звуков

Очевидно, что включение в описанный подход в качестве параметров для идентификации некоторых статистических показателей речи (распределение частоты основного тона и т.д.) позволит несколько уменьшить вероятность ошибки.

Глава 3. АЛГОРИТМЫ РАСПОЗНАВАНИЯ РЕЧИ

Автоматические системы распознавания речи (ACPP) могут быть классифицированы на основе нескольких видов признаков: тип речи, который должна распознавать система; множество дикторов, по отношению к которому необходимо обеспечить устойчивое распознавание; объем словаря, который необходим распознавать; полнота словаря, используемого для описания речевого сообщения

Среди типов речи можно выделить дискретную и слитную речь. Дискретная речь определяется как режим произношения слов, при котором паузы между словами значительно больше, чем внутрисловные паузы при нормальном темпе речи. Обычно принимается, что пауза между словами должна быть не менее 0,5 с. Слитная речь соответствует нашему произвольному режиму произношения слов.

ACPP (по признаку множества дикторов для устойчивого распознавания) подразделяют на зависимые от индивидуальности голоса диктора (дикторозависимые) и независимые (дикторонезависимые). Это разделение не означает, что дикторозависимая система будет распознавать слова сказанные голосом только избранного пользователя, а всех прочих — нет; скорее — точность распознавания слов «своего» пользователя будет больше, чем средняя точность, взятая по всем «чужим» пользователям. В соответствии со сказанным, термин дикторонезависимость означает, что точность для определенного пользователя близка к средней точности, взятой по всем возможным дикторам.

ACPP по объему словаря можно разделить на две категории: системы с малыми и большими словарями. Эти две категории подразумевают применение принципиально различных методов обучения. Если словарь мал, то имеется возможность провести прямое обучение, т.е. пользователь системы может непосредственно диктовать требуемые для распознавания слова. Для большого словаря возможности продиктовать системе все слова, содержащиеся в словаре, нет, так как это заняло бы много часов

или даже суток. Идея обучения такого рода систем состоит в синтезе акустических признаков слова из признаков последовательности более мелких единиц речи (фонем, трифонов, пентелефонов и т.д.)

Полнота словаря подразумевает, что всякое слово речевого сообщения содержится в словаре системы. Однако часто возникает практическая потребность в решении задач, когда словарь системы неполон. Например, в задаче выделения ключевых слов из потока слитной речи необходимо искать в речевом сообщении слова и фразы, которые не составляют полного словаря.

Таким образом, приведенную классификацию, можно представить в виде таблицы (табл. 3.1).

Таблица 3.1
Классификационные признаки АСРР

Признаки	Объем словаря	Полнота словаря	Тип речи	Индивидуальность голоса
1	Малый (<100)	Полный	Дискретная	Дикторозависимость
2	Большой	Неполный	Слитная	Дикторонезависимость

Безусловно можно провести более тонкую классификацию АСРР, используя в качестве классификационных признаков типы каналов передачи речи, которые специфическим образом искажают речь и вызывают необходимость дополнять АСРР методами адаптации под тип канала и т.п., но здесь мы не будем этого делать, поскольку такая классификация не требует разработки новых моделей распознавания.

Итак, цель АСРР, основанных на статистических методах, заключается в вычислении наиболее вероятной последовательности слов \hat{W} , которая генерирует последовательность наблюдений H при заданных параметрах модели λ , т.е.

$$\hat{W} = \arg \max_W P(W)P(H | W, \lambda), \quad (3.1)$$

где $P(W)$ — априорная вероятность последовательности слов W .

В работе [8] описана функциональная структура АСРР, позволяющая вычислять значение выражения (3.1) (рис. 3.1). Не-

смотря на то, что эта работа была написана почти 30 лет назад, вряд ли мы можем дополнить эту схему какими-либо принципиально новыми блоками.

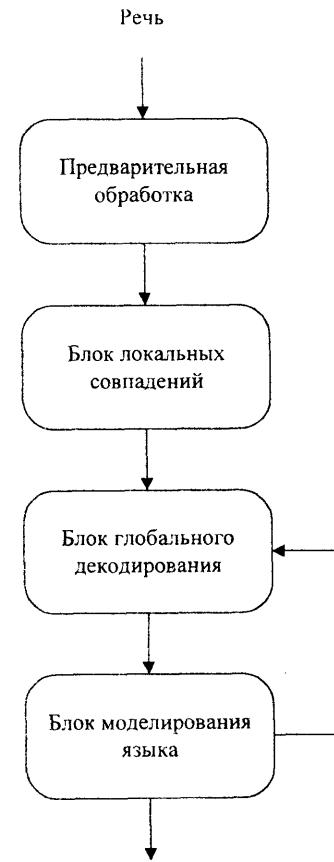


Рис. 3.1. Функциональная структура АСРР

Блок предварительной обработки кодирует речевое высказывание последовательностью векторов наблюдений и передает эту последовательность в блок локальных совпадений.

В передаваемой последовательности есть подпоследовательности относительно похожих состояний. Задача блока локальных совпадений — выявить эти подпоследовательности и про-

Таблица 3.2

Направления развития СММ

Наблюдение	Состояние	Вероятность $b_j(\mathbf{h}_t)$	Размерность
Мел спектр $\mathbf{h}_t \in R^n, \forall t$	Фонема	Одномодальное распределение Гаусса	Иерархические СММ
Мел кепстр $\mathbf{h}_t \in R^n, \forall t$	Трифон	Гауссовые смеси	—
Мел коэффициенты линейного предсказания $\mathbf{h}_t \in R^n, \forall t$	Пентафон	Смесь нормальных распределений с линейной авторегрессией	—
Характеристики формант $\mathbf{h}_t \in R^{n(1)}$	—	Смесь нормальных распределений с квадратичной авторегрессией	—

варьируется в процессе речеобразования от фонеме к фонеме и от диктора к диктору).

Известно, что с одной стороны, звучание той или иной фонемы зависит от ее окружения или (иными словами) от контекста [58,59], в котором фонема сказана, с другой стороны, особенностью цепей Маркова является возможность задать вероятностную зависимость только между парой соседних состояний. Для согласования этих двух положений необходимо определять состояния модели таким образом, чтобы в него входила фонема со своим окружением. На этом пути возникли состояния, названные трифонами и пентафонами [60] (табл. 3.2).

В определении СММ (приложение 1) говорится о допущении, что вероятность текущего наблюдения \mathbf{h}_t зависит только от текущего скрытого состояния и не зависит от других переменных: $P(\mathbf{h}_t | \mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{T-1}, q_0, q_1, \dots, q_{T-1}) = p(\mathbf{h}_t | q_t)$. В рамках этого допущения были развиты одномодальные распределения Гаусса и гауссовые смеси [61], параметры которых вычисляются с помощью ЕМ-алгоритма (приложение 2).

В работе [62], эта условная зависимость наблюдения \mathbf{h}_t расширяется положением, что наблюдение зависит не только от текущего скрытого состояния, но и от наблюдаемого окруже-

вести их сверку. В некоторых модификациях АСРР этим блоком пренебрегают.

Глобальный декодер классифицирует последовательность различных состояний, поступившую с выхода блока локальных совпадений. Эта классификация заключается в том, что в рамках кодовой книги, в которой содержатся эталонные последовательности, ищутся несколько последовательностей, наиболее сходных (в вероятностном смысле) с входной последовательностью.

Последний блок рисунка 3.1. представляет собой модель языка, с помощью которой проводится дальнейший анализ найденных эталонных последовательностей с точки зрения лексики. Языковая модель приписывает этим последовательностям некоторые веса, в соответствии с которыми можно принять окончательное решение о наиболее вероятной последовательности слов, описывающей входную последовательность звуков.

Использование СММ в качестве модели блока глобального декодирования является доминирующим и в случае дискретной, и в случае непрерывной речи, поэтому здесь является уместным изложить направления развития этих моделей, с которыми связывается перспектива в области распознавания речи.

По сути дела всякое направление развития СММ связано с входящими в ее определение (см. приложение 1) параметрами. Это хорошо видно на примере таблицы 3.2.

Исследование представлений вектора наблюдения случайного процесса \mathbf{h}_t привели к двум существенно различным классам векторов (табл. 3.2). Первый класс связан с векторами, размерность которых сохраняется во времени. К этому классу относятся рассмотренные в первой главе мел кепстральные коэффициенты [56] и мел коэффициенты линейного предсказания [57]. Ко второму классу относятся вектора, размерность которых не сохраняется с течением времени. В приложении 3 рассмотрено формантное представление [30] речевого сигнала, относящиеся к этому классу (количество формант

ния, т.е. $P(\mathbf{h}_t | \mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{t-1}, q_0, q_1, \dots, q_{T-1}) = p(\mathbf{h}_t | \mathbf{h}_{t-\kappa}, \dots, \mathbf{h}_{t+n}, q_t)$. Правую часть этого выражения аппроксимируют различными формами нормального процесса с авторегрессией (табл. 3.2), которые будут описаны ниже.

Четвертый столбец таблицы 3.2 показывает исследования возможности ввести в СММ модель языка, которая отражает зависимости в чередованиях слов. Необходимо отметить, что оперировать такими зависимостями можно только в случае, если мы обладаем достаточно полным словарем.

В столбце «разное» указаны модели, которые связаны с введением неоднородности в СММ, например ИКДП-подход Винценко [9, 63] (эта модель начала развиваться несколько раньше, чем сама СММ) и с построением иерархических моделей, каждый из слоев которой представляет СММ. Эта модель будет обсуждаться ниже (см. разд. 3.5).

Методы оценки параметров модели нельзя отнести непосредственно к СММ, поэтому они не вошли в таблицу 3.2. Однако на этом пути были развиты следующие методы: максимального правдоподобия, на основе которого был разработан ЕМ-алгоритм [21], оценки значения вероятности $b_j(\mathbf{h}_i)$ нейронными сетями [15], максимизации относительной информации [15], которые позволяют учитывать относительные местоположения классов в пространстве признаков и таким образом более точно моделировать случайный процесс.

3.1. СММ-модели фонем, трифонов и пентафонов

В фонетике минимальную звуковую единицу речи принято называть фонемой, а каждый вариант звучания фонемы, который зависит от контекста и интонации фразы, называют аллофоном (в соответствии с дескриптивной фонетической школой [59]). Соответствующее фонеме акустическое представление (фон) не является однородным на всем своем протяжении, а обладает внутренней структурой. Для описания этой неоднородности фонему моделируют как последовательность состояний. Каждая

фонема рассматривается как СММ-модель, определенная на своем множестве состояний связанными переходами. СММ-модели фонем имеют обычно три состояния (2,3,4) и связи (рис. 3.2).

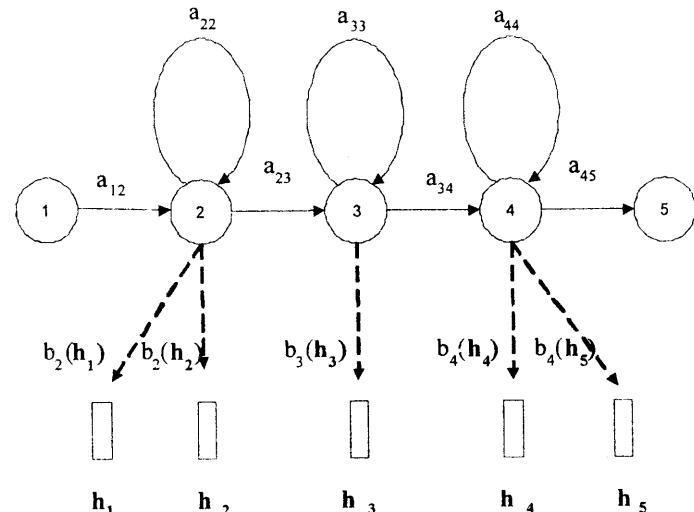


Рис. 3.2. СММ-модель фонемы

Начальное и конечное состояния (1, 5) добавлены для объединения моделей различных фонем в последовательности, которые образуют слова и высказывания.

СММ-модель фонемы представляет генератор последовательности векторов признаков, т.е. это вероятностная машина с конечным числом состояний, которая каждый момент времени t переходит из состояния в состояние, излучая при этом вектор признаков \mathbf{h}_t в соответствии со значениями условных вероятностей $b_j(\mathbf{h}_t)$. Переход из состояния i в состояние j также подчиняется вероятностному закону и определяется матрицей вероятностей a_{ij} (приложение 1). На рисунке 3.2. показан пример такого процесса, когда система проходит последовательность состояний $Q = \{1, 2, 2, 3, 4, 4, 5\}$ и генерирует последовательность векторов наблюдений $H = \{\mathbf{h}_1, \dots, \mathbf{h}_5\}$.

Как уже говорилось, контекст очень сильно влияет на звучание той или иной фонемы, и для достижения точности при распознавании необходимо учитывать эту контекстную зависимость, т.е. одну и ту же фонему в различных окружениях необходимо описывать различными СММ-моделями. Для решения этой задачи используются трифоны, где каждая фонема и пара из левого и правого соседа имеет отдельную СММ-модель. Например, запись $x-y+z$ представляет фонему y , перед которой стоит фонема x , после которой следует фонема z . Например, английская фраза: «Beat it» имеет фонетическое представление — *sil b iy t ih t sil*. Символ «*sil*» означает паузу. Трифонное представление этой фразы имеет вид:

sil sil-b+iy b-iy+t iy-t+ih t-ih+t ih-t+sil sil .

Заметим, что в этой фразе трифонный контекст пересекает границы слова, и одна и та же фонема t в разных контекстах представляется разными СММ-моделями. Трифоны $iy-t+ih$, $t-ih+t$ это так называемые межсловные трифоны, использование которых повышает точность распознавания по сравнению с распознаванием на основе представлений вида

sil sil-b+iy b-iy+t iy-t+sil sil-ih+t ih-t+sil sil .

Представление условных вероятностей $b(\mathbf{h})$ в виде гауссовых смесей (приложение 2) позволяет достаточно точно описывать каждое состояние СММ-модели. Однако при использовании трифонов мы имеем систему, в которой необходимо вычислять очень большое количество параметров. Например, система, включающая в себя межсловные трифоны, имеет их приблизительно 60 000. На практике десятикомпонентная гауссова смесь дает хорошее качество распознавания. Предполагая ковариационную матрицу диагональной и вектор признаков, по крайней мере, 40-мерным, получим, что на одно состояние приходится 790 параметров, которые необходимо оценить. Следовательно, 60 000 моделей, состоящих из трех состояний, требуют вычисления приблизительно 142 млн параметров.

Для достоверной оценки такой модели необходимо либо иметь представительную выборку обучающего материала, либо ис-

кать пути сокращения размерности модели. Путь сокращения размерности модели состоит в том, чтобы связать вместе состояния, которые акустически неразличимы [64–70], т.е. разнообразие таких неразличимых состояний заменить одним состоянием. Эта операция проиллюстрирована на рисунке 3.3. В верхней части рисунка каждый трифон имеет свое собственное распределение. После связывания некоторые состояния используют общее распределение.

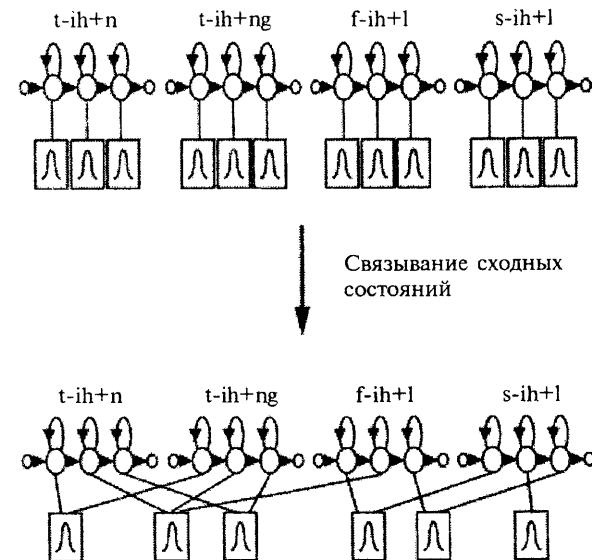


Рис. 3.3. Процедура связывания состояний

Для решения вопроса, какие состояния связываются, используется фонетическое дерево решений [71, 72]. Такое дерево создается для каждого связанного состояния, и оно является бинарным, т.е. в каждом узле дается ответ «да» или «нет» на некоторый фонетический вопрос. Например, «является ли левый контекст назальным звуком». В начале все связываемые состояния находятся в корневом узле дерева. По мере движения по дереву состояния расщепляются в зависимости от получаемых ответов и это продолжается, пока состояния не достигнут

конечных узлов дерева. Далее все состояния, оказавшиеся в одном конечном узле, связываются. Например, рисунок 3.4 иллюстрирует связывание центральных состояний всех трифонов фонемы *aw*. Опускаясь по дереву решений в соответствии с ответами на фонетические вопросы, состояния оказываются в конечных узлах дерева, закрашенных серым цветом. Например, центральное состояние трифона *s—aw+n* оказывается во втором справа конечном узле, поскольку его правый контекст является переднеязычным назальным и не является взрывным звуком.

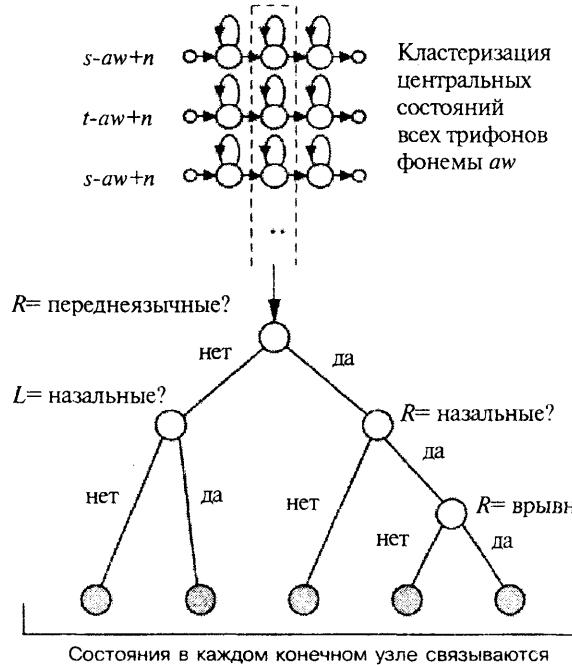


Рис. 3.4. Кластеризация с использованием бинарного дерева решений

Фонетические вопросы в каждом узле дерева выбираются в соответствии с принципом максимума вероятности обучающих данных оказаться кластеризованными таким образом. На практике кластеризация, основанная на фонетическом дереве решений, дает набор кластеров, которые имеют достаточное коли-

чество обучающих данных для оценки параметров выходной функции распределения. Более того, сконструированные деревья могут быть использованы для синтеза моделей для любого контекста, независимо от того, был он представлен в обучении или нет. Наконец, кластеризация, основанная на фонетическом дереве решений, позволяет увеличить контекст и использовать не трифоны, а пентафоны.

Пентафон — это структура во всем аналогичная трифону. Для ее построения могут использоваться от трех до шести внутренних, входное и выходное состояния. Однако топология связей внутри пентафона более разнообразна чем в трифоне (рис. 3.5). Структура, показанная на рисунке 3.5, *a*, допускает переходы только из состояния самого в себя, либо в состояние на один номер выше; на рисунке 3.5, *б* структура допускает переходы из состояния в состояние на два номера выше; на рисунке 3.5, *в* показано, что случайный процесс, соответствующий трифону, может развиваться по двум взаимоисключающим путям.

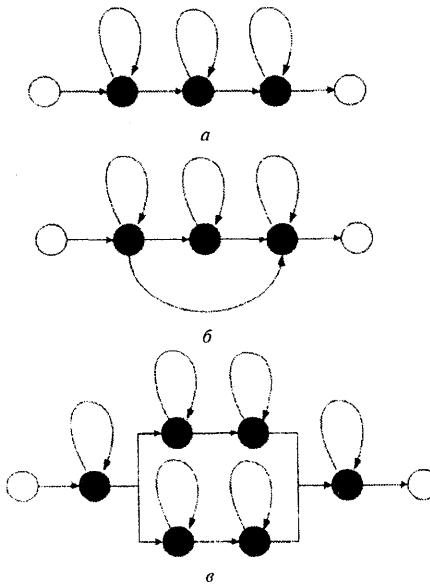


Рис. 3.5. Внутренняя структура пентафона

3.2. Использование нормального распределения с линейной авторегрессией для аппроксимации апостериорной вероятности

Как уже ранее упоминалось, наиболее развитой моделью аппроксимации апостериорной вероятности $b_j(\mathbf{h}_t)$ являются гауссовые смеси, параметры которых вычисляются в течение обучения на основе ЕМ-алгоритма, приведенного в приложении 2. Однако такое представление этой вероятности не является единственным. Здесь мы приведем примеры использования смесей нормальных распределений с линейной и квадратичной авторегрессией.

В разделе 3.1. был показан способ введения контекстной зависимости на фонетическом уровне с помощью использования трифонов и пентафонов. Очевидно, что контекстную зависимость можно ввести и на акустическом уровне. В основе акустической контекстной зависимости лежат два положения:

- процесс изменения положения артикуляционных органов является гладким, а не скачкообразным;
- процесс перестройки положения артикуляционных органов несет в себе информацию о конечном состоянии, в которое им необходимо перестроиться.

Рассмотрим случай, когда компоненты вектора признаков можно считать независимыми случайными процессами и справедливо

$$b_j(\mathbf{h}_t) = \prod_{i=0}^n b_j(h_{it}).$$

Аппроксимируем покомпонентную апостериорную плотность вероятности $b_j(h_{it})$ нормальным процессом авторегрессии

$$b_j(h_{kt}) = \frac{1}{\sqrt{2\pi}\sigma_{kj}} \exp \left\{ -\frac{1}{2\sigma_{kj}^2} \left(h_{kt} - g_{0kj} - \sum_{i=1}^n g_{kij} h_{kt-i} - \sum_{i=1}^m v_{kij} h_{kt+i} \right)^2 \right\}, \quad (3.2)$$

где n — глубина авторегрессии. В дальнейшем индекс, соответствующий номеру компоненты, будем опускать, поскольку все вычисления однотипны для каждой компоненты. Использова-

ние такого типа авторегрессии не является физически реализуемым, поскольку для предсказания значения точки в момент времени t используются значения в будущие моменты времени. В вычислительных системах, обладающих памятью, мы можем позволить себе эмулировать такую систему, проигрывая лишь в том, что отстаем в принятии решения от реального времени на m тактов.

Основная задача состоит в оценке параметров распределения (3.2). Ее можно решить комбинируя:

- метод максимального правдоподобия, подробно описанного для этого случая в работе [17];
- алгоритм прямого обратного хода (приложение 1.1).

Метод максимального правдоподобия требует, чтобы для всех векторов наблюдения, достоверно принадлежащих к состоянию с номером j , выполнялось условие

$$F = \sum_{t=0}^{T-1} \ln b_j(h_t) \rightarrow \min.$$

Если же известна априорная вероятность принадлежности вектора x , состоянию $j = p_t^j$, то условие правдоподобия можно записать в виде

$$F = \sum_{t=0}^{T-1} p_t^j \ln b_j(h_t) \rightarrow \min. \quad (3.3)$$

Раскроем выражение (3.3)

$$F = \sum_{t=1}^T p_t^j \left(-\frac{1}{2} \ln 2\pi - \ln \sigma_j - \frac{1}{2\sigma_j^2} \left(h_t - g_{0j} - \sum_{i=1}^n g_{ij} h_{t-i} - \sum_{i=1}^m v_{ij} h_{t+i} \right)^2 \right)$$

и продифференцируем по неизвестным параметрам g , v . В результате получим систему линейных уравнений

$$\begin{aligned} g_{0j} \sum_{t=1}^T p_t^j - \sum_{i=1}^n g_{ij} \sum_{t=1}^T p_t^j h_{t-i} - \sum_{i=1}^m v_{ij} \sum_{t=1}^T p_t^j h_{t+i} &= \sum_{t=1}^T p_t^j h_t; \\ g_{0j} \sum_{t=1}^T p_t^j h_{t-s} - \sum_{i=1}^n g_{ij} \sum_{t=1}^T p_t^j h_{t-i} h_{t-s} - \sum_{i=1}^m v_{ij} \sum_{t=1}^T p_t^j h_{t+i} h_{t-s} &= \sum_{t=1}^T p_t^j h_t h_{t-s}, \end{aligned}$$

$s = 1, \dots, n;$

$$g_{0j} \sum_{t=1}^T p_t^j h_{t+s} - \sum_{i=1}^n g_{ij} \sum_{t=1}^T p_t^j h_{t-i} h_{t+s} - \sum_{i=1}^m v_{ij} \sum_{t=1}^T p_t^j h_{t+i} h_{t+s} = \sum_{t=1}^T p_t^j h_t h_{t+s},$$

$$s = 1, \dots, m;$$

$$\sigma_j = \frac{\sum_{t=1}^T p_t^j \left(h_t - g_{0j} - \sum_{i=1}^n g_{ij} h_{t-i} - \sum_{i=1}^m v_{ij} h_{t+i} \right)^2}{\sum_{t=1}^T p_t^j}. \quad (3.4)$$

Значения априорных вероятностей являются неизвестными и для их начальной инициализации необходимо воспользоваться дополнительными предположениями:

1) транскрипция фразы, на основе которой проводится вычисление параметров (обучение), известна (это допущение достаточно общее и используется для систем распознавания с большими словарями);

2) разбить речевое сообщение на участки пропорционально известным средним длительностям фонем и положить на этих участках

$$p_t^{j(0)} = \begin{cases} 1, & j = 9, \\ 0, & j \neq 9, \end{cases}, \quad (3.5)$$

где 9 — символ транскрипции соответствующий сегменту, который включает в себя момент времени t . Если данных о средних длительностях фонем нет, то можно допустить, что все фонемы имеют одинаковую длительность.

Подстановка значений (3.5) в систему линейных уравнений позволит вычислить нулевое приближение коэффициентов линейной авторегрессии и, как следствие, нулевое приближение плотности вероятностей $b_j^{(0)}(h_t)$.

Параллельно с вычислениями приближений нормального процесса с авторегрессией необходимо вычислять приближения матрицы переходных вероятностей между состояниями. Этот итерационный процесс задается алгоритмом Баума—Уолша (приложение 1) при начальном допущении, что все возможные

переходы их состояния равновероятны, поэтому мы не будем здесь его повторять. Следует обратить внимание, что под словами «всевозможные переходы» имеется ввиду схема соединения состояний. Если, например, использована схема, показанная на рисунке 3.5, *a*, то возможных переходов из состояния в состояние всего два: сам в себя и в следующего соседа. Использование схемы переходов, показанной на рисунке 3.5, *б*, приведет к необходимости описывать три возможных перехода из состояния в состояние.

Подстановка r -го приближения величин $b_j^{(r)}(h_t)$ и $a_{ij}^{(r)}$ в переменные прямого обратного хода позволяют вычислить их соответствующее приближение

$$\alpha_j^{(r)}(1) = c_j, \quad \alpha_j^{(r)}(t) = b_j^{(r)}(h_t) \sum_{i=1}^N a_{ij}^{(r)}(t-1), \quad \beta_j^{(r)}(t) = \sum_{i=1}^N a_{ji}^{(r)} b_i^{(r)}(h_{t+1}) \beta_i^{(r)}(t+1),$$

подстановка которого в выражение

$$p_t^{j(r)} = \frac{\alpha_j^{(r)}(t) \beta_j^{(r)}(t)}{\alpha_N^{(r)}(T)},$$

позволяет найти новое приближение априорной вероятности отсчета. Далее, продолжая рекурсивные вычисления, решим систему линейных уравнений (3.4) и определим новые значения коэффициентов авторегрессии.

В работе [62] было описано использование смеси авторегрессионных процессов для описания состояния, если вектор наблюдений образован 13-ю мелкоэффициентами линейного предсказания. Однако вводилось ограничение, что ненулевым коэффициентом является коэффициент с глубиной 5. Проведенные экспериментальные исследования показали, что по отношению к гауссовой смеси использование авторегрессии приводит к увеличению точности распознавания примерно на 3 % для одной, трех и шести компонент в смесях. Авторы [62] указывают на практические трудности при использовании боль-

шего числа параметров авторегрессии, но не объясняют характера этих трудностей.

Авторами данной работы проводились экспериментальные исследования по распознаванию дискретной речи с одной компонентой смеси и различными значениями глубин авторегрессии (табл. 3.3). База состояла из 1 000 тестирующих слов, записанных при частоте оцифровки 8 кГц и 16 Бт на отсчет. В качестве векторов наблюдений использовались 24-мерные мел спектры при длительности фрейма 25 мс и его смещении 10 мс.

Таблица 3.3

Экспериментальные результаты точности распознавания в зависимости от глубины авторегрессии

Глубина запаздывающей авторегрессии	Глубина опережающей авторегрессии	Точность распознавания, %
0	0	82,28
0	6	20,42
3	0	83,40
6	0	90,49
9	0	94,59
7	3	62,66

На основании полученных данных можно сделать вывод, что повышение глубины запаздывающей части авторегрессии повышает точность распознавания. Здесь мы ограничились глубиной 9, поскольку дальнейшее увеличение глубины приводит к техническим затруднениям, связанным с переходом за границу записи. Всякая попытка использовать коэффициенты опережающей авторегрессии приводила к значительной потере точности.

3.3. Использование нормальных распределений с квадратичной регрессией для аппроксимации апостериорной вероятности

Основная гипотеза, которую используют при построении систем распознавания речи, связана с тем, что в пространстве рече-

вых параметров для каждой фонемы можно выделить присущую только ей область [8]. Траектория речевых параметров, возникающая в результате произношения того или иного слова, должна пройти через области, соответствующие последовательности элементов (фонем, трифонов, пентафонов) данного слова.

Практика исследования систем распознавания речи и динамики органов артикуляции [45] позволяет утверждать, что благодаря инерции этих органов, даже при нормальном темпе произношения речи, им не удается занять положение, соответствующее произносимой фонеме. Органы артикуляции лишь стремятся занять требуемое положение. Такая форма динамики вокального тракта позволяет предположить, что каждому речевому элементу соответствует некоторая виртуальная точка в пространстве параметров (т.е. точка, лежащая в пространстве параметров, но удаленная от области, в которой проходит траектория процесса). Для выделения того или иного речевого элемента важна не столько близость к виртуальной точке, сколько тип движения относительно нее (рис. 3.6).

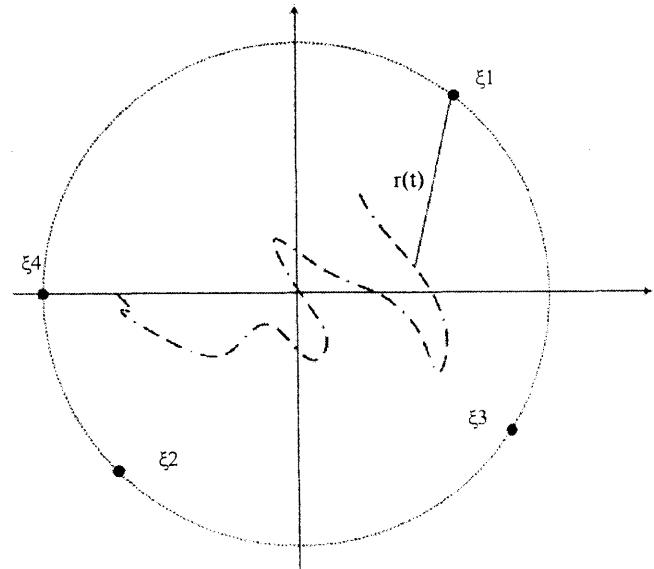


Рис. 3.6. Способ представления виртуальных точек

Зададим плотность вероятности в виде

$$b_k(\mathbf{h}_t) = \frac{1}{\sqrt{2\pi \det \Psi}} \exp \left\{ -\sum_{i,j=1}^{m,m} \psi_{ij}^{-1} r(\mathbf{h}_{t-i}, \xi_k) r(\mathbf{h}_{t-i}, \xi_k) \right\},$$

где $r(x, y)$ — евклидово расстояние; ψ_{ij}^{-1} — элементы обратной матрицы.

Рассмотрим случай, когда матрица Ψ диагональная. Введем обозначения $\alpha_i = \psi_{ii}$ и запишем функционал, который используется в методе максимального правдоподобия для вычисления коэффициентов матрицы и компонент векторов $\xi_k \in R^n \quad \forall k$

$$F_\xi = \sum_{t=0}^T R_{t,\xi} = \sum_{t=0}^{T-1} \sum_{i=1}^m \left(\frac{1}{2} \ln \alpha_i + \alpha_i r^2(\mathbf{h}_{t-i}, \xi) \right), \quad (3.6)$$

где T — длительность выборки обучающего случайного процесса; m — глубина регрессии.

Минимизируем (3.6) при условии, что всякий вектор лежит на гиперсфере радиуса D , т.е. что $D^2 = \sum_{j=1}^n \xi_j^2 = \text{const} \quad \forall k$, методом Лагранжа. Поскольку все расчеты аналогичны для любого вектора ξ_k , опустим индекс k .

Опустим громоздкие выкладки, связанные с дифференцированием функционала вида

$$F_L = F_\xi + \lambda \left(D^2 - \sum_{j=1}^n \xi_j^2 \right), \quad (3.7)$$

по компонентам вектора ξ и по неопределенному множителю λ . Для компонент вектора ξ получим

$$\xi_k = \frac{1}{\lambda - mT} \sum_{i=1}^m \alpha_i X_{k,i}, \quad (3.8)$$

где

$$\lambda = \sqrt{\frac{\sum_{k=1}^n \left(\sum_{i=1}^m \alpha_i X_{k,i} \right)^2}{D^2}}; \quad X_{k,i} = \sum_{t=0}^{T-1} h_{k,t-i}.$$

Дифференцируя (3.7) по весовым коэффициентам α_i , получим

$$\sum_{k=0}^N X_{k,i} \xi_k = \frac{1}{2} (\Phi_i + D^2), \quad (3.9)$$

где введено обозначения $\Phi_i = \sum_{k=1}^n \sum_{t=0}^{T-1} h_{k,t-i}^2$.

Подстановка (3.8) в (3.9) приводит к нелинейной системе алгебраических уравнений относительно весовых коэффициентов

$$\sum_{i,j=1}^m \left(\frac{A_{ij}}{D^2} - \frac{4A_{ni}A_{nj}}{(\Phi_n + D^2)^2} \right) \alpha_i \alpha_j + \frac{4mT}{\Phi_n + D^2} \sum_{j=1}^m A_{nj} \alpha_j - m^2 T^2 = 0, \quad (3.10)$$

которая может быть решена только численно. Здесь введено обозначение $A_{ij} = \sum_{k=1}^n X_{k,i} X_{k,j}$. Из всех 2^m возможных решений системы (3.10) мы должны отобрать лишь те, где все α_i действительные, $\prod \alpha_i > 0$, и $\sum \alpha_i > 0$.

Эксперименты, проведенные при тех же условиях на речевые сообщения, как и в случае п.3.2, показывают, что введение виртуальных точек и глубины авторегрессии 6 повышают вероятность распознавания слов от 1,5 до 4 % (по отношению к моделированию состояния гауссовой смесью из трех компонент), но только при удачном выборе значения нормы D для каждого диктора.

3.4. Модель языка

В работе [58] для модели языка выделены следующие уровни: фонетический, фонологический, морфологический, лексический, синтаксический и семантический. Все уровни несут информацию о структуре разговорного языка, которая увеличивает шансы принятия верного решения о произнесенной фразе. На фонетическом уровне происходит преобразование наблюдае-

мой акустической волны в цепочки фонем. На фонологическом уровне накладываются ограничения на возможные последовательности фонем. Эти ограничения основываются на полезной априорной информации о вероятности следования фонем друг за другом. Далее, на морфологическом уровне оперируют со слогоподобными единицами речи (морфемами). Они накладывают ограничение уже на структуру слова, подчиняясь закономерностям моделируемого естественного языка. Лексический уровень охватывает слова и словоформы того или иного естественного языка, т.е. устанавливает словарь языка и вероятностные связи в последовательностях слов. Семантика устанавливает соотношения между объектами действительности и словами, их обозначающими. Она является высшим уровнем языка.

Необходимо отметить, что математические модели АСРР упрощают структуру языка, ограничиваясь лишь самыми простейшими вариантами. Как правило, игнорируют морфологический и семантический уровни.

В практических приложениях цель применения модели языка состоит в создании механизма расчета априорной вероятности сообщения из формулы (3.1), где W — последовательность слов $\{w_1, w_2, \dots, w_m\}$. Применение N -грамм моделей успешно реализует эту цель [71]. Принцип построения N -грамм состоит в расчете вероятности последовательности слов, причем учитываются условные зависимости в цепочках слов длиной до N включительно.

$$P(W = \{w_1, w_2, \dots, w_m\}) = \prod_{i=1}^m p(w_i | w_{i-1}, \dots, w_{i-N+1}).$$

Оценка условных вероятностей проводится на основе анализа текстовых сообщений и подсчете количества встреч последовательности слов $C(w_i, w_{i-1}, \dots, w_{i-N+1})$:

$$p(w_i | w_{i-1}, \dots, w_{i-N+1}) = \frac{C(w_i, w_{i-1}, \dots, w_{i-N+1})}{C(w_{i-1}, w_{i-2}, \dots, w_{i-N+1})}.$$

Зачастую используется нулевое приближение N -граммой модели (не учитывается зависимость следования слов в сообще-

нии), т.е. предполагается, что появление слова есть событие, независимое от предшествующих слов. Таким образом, вероятность сообщения $P(W)$ рассчитывается в виде

$$P(W) = \prod_{i=1}^N p(w_i),$$

где $p(w_i)$ — априорная вероятность появления слова w_i , которую так же следует определять на основе анализа текстовой информации.

Например, в задаче поиска ключевых слов используется нулевое приближение N -грамм, что позволяет применять циклическую сеть распознавания (рис. 3.7). Она отличается тем, что позволяет любые комбинации слов из словаря. Такая сеть наиболее приемлема в данной задаче, поскольку нет никакой априорной информации о возможных закономерностях следования слов при произвольно заданном малом словаре.

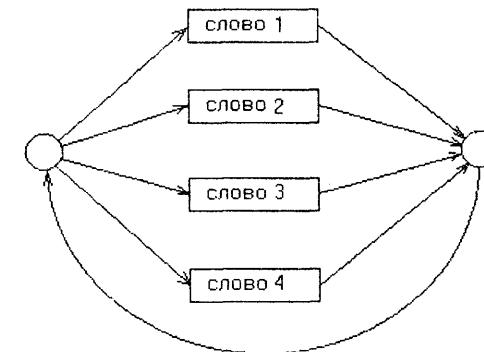


Рис. 3.7. Циклическая сеть для системы со словарем из 4 слов

3.5. Иерархические СММ-модели

Различные способы определения речевых элементов для СММ (фонемы, трифоны и т.д.) требуют искусственно вводить величину контекстной зависимости, с помощью числа дополнительных фонем справа и слева. Обратим внимание

на то, что при таких условиях учета контекста мы сталкиваемся с двумя очевидными трудностями: 1) создаем речевые элементы, которые не встречаются в языке; 2) точно не знаем, когда остановиться в расширении границ контекстной зависимости.

Рассмотрим структуру иерархической системы, на примере двух слоев (рис. 3.8), подразумевая, что принцип соединения элементов может быть распространен на произвольное количество слоев. Элементы первого слоя содержат информацию о распределениях вероятностей компонент вектора признаков в различных состояниях речи и позволяют вычислять вероятности соответствия поступающего вектора признаков состоянию. Под состояниями речи можно подразумевать устойчивое во времени поведение компонент вектора признаков.

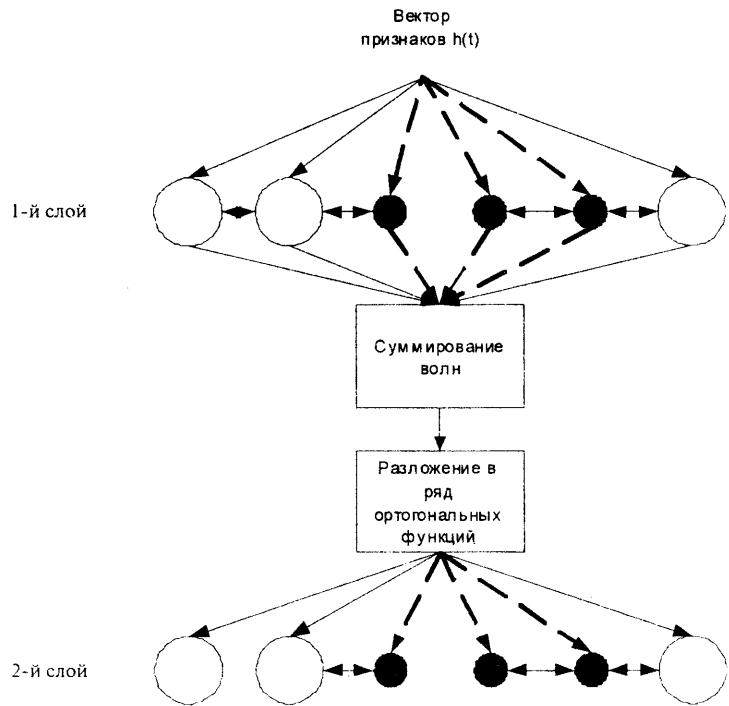


Рис. 3.8. Структура иерархической системы распознавания

В момент времени поступления на первый слой вектора признаков, каждый из элементов вычисляет значение условной вероятности $b_i^{(1)}(t) = p(s_i^{(1)} | \mathbf{h}_t)$, где $s_i^{(1)}$ — i -ый элемент первого слоя, $i = 1, \dots, M^{(1)}$. Каждый из элементов слоя обладает памятью и способен сохранять последовательность из K значений вычисленных вероятностей $b_i^{(1)}(t)$. В момент времени, когда элементом накоплено K значений вероятностей, на своем выходе он генерирует синусоидальную волну, которая имеет вид

$$\psi_i(t) = \sum_{j=1}^K b_i^{(1)}(j) \delta(T(j-1) \leq t < jT) \sin\left(\pi \frac{ij}{T} t\right), \quad (3.11)$$

где функция $\delta(\cdot)$ определена как

$$\delta(T_1 \leq t < T_2) = \begin{cases} 1, & \text{если } \text{правда} \\ 0, & \text{иначе} \end{cases},$$

T — период времени между появлениями двух последовательных векторов наблюдения.

Формула (3.11) означает, что выходные сигналы элементов слоя, во-первых, кратны периоду наблюдения T и, во-вторых, модулированы значениями вероятностей $b_i^{(1)}(t)$.

Между слоями находится элемент, который суммирует все волны элементов первого слоя

$$Q(t) = \sum_{i=1}^{M^{(1)}} \psi_i(t). \quad (3.12)$$

Затем выходная волна $Q(t)$ этого элемента разлагается в ряд Фурье по периоду КТ. Спектр этой волны S , является вектором признаков для элементов второго слоя. Использование спектра волны в качестве вектора признаков является лишь частным случаем, в общем случае это может быть любая целесообразная спектральная оценка, которая сокращает размерность представления данных.

На следующем такте работы элементов первого слоя D новых значений вероятностей $b_i^{(1)}(t)$ замещают D старых значений, как это показано в таблице 3.4, и процедура генерации

волн, их суммирования и разложения в спектр Фурье повторяется каждые D тактов.

Таблица 3.4

Пример замещения данных в элементах слоев иерархической системы

Содержание памяти i -го элемента первого слоя в первый такт работы ($K = 5$)	Содержание памяти i -го элемента первого слоя во второй такт работы ($K = 5, D = 2$)
$b_i^{(1)}(1)$	$b_i^{(1)}(3)$
$b_i^{(1)}(2)$	$b_i^{(1)}(3)$
$b_i^{(1)}(3)$	$b_i^{(1)}(5)$
$b_i^{(1)}(4)$	$b_i^{(1)}(6)$
$b_i^{(1)}(5)$	$b_i^{(1)}(7)$

Каждый элемент второго слоя состоит из элементов, которые характеризуют состояния процесса. Под состояниями второго слоя можно понимать фонемы или сочетания фонем. Максимальное количество фонем, участвующих в сочетании, зависит от отношения средней длительности фонемы к величине смещения D умноженной на длительность окна анализа T .

Операции (3.11), (3.12) и спектральное оценивание можно распространить на более глубокие слои; при этом будет расти длительность контекстной зависимости как $D^{n-1}T$, где n — номер слоя.

Заметим, что элементы каждого из введенных слоев можно использовать в качестве множества для СММ и решать задачи, описанные в приложении 1. Причем, каждый слой способен принимать решение о том, какая фраза была произнесена.

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ 1

1. Описание однородной СММ и связанных с ней задач

В области обработки сигналов часто возникают задачи, связанные с исследованием недоступных для непосредственного наблюдения физических систем через некоторые сопутствующие процессы, которые могут наблюдаться непосредственно.

Целью настоящего исследования является изучение поведения физической системы как источника сигнала в виде последовательности элементарных событий. Можно привести достаточно широкий класс примеров такого рода систем — генераторов наблюдаемого нестационарного случайного процесса, который образуется при прохождении системой скрытой для наблюдателя последовательности элементарных событий. Это может быть: наблюдение за поведением автомобиля по измерениям акустических параметров работы двигателя и неизвестных текущих состояниях коробки передач; процессы речеобразования, где в качестве наблюдаемых параметров выступают акустические признаки, а в качестве скрытых — понятия, которые говорящий желает выразить; процессы динамики земной коры, которые характеризуются наблюдаемыми волнами сейсмической активности и скрытыми состояниями — глубиной источников этих волн. Для описания таких процессов используются однородные СММ. Поскольку в рамках данной работы используются только однородные СММ, то термин «однородный» мы опустим.

Введем обозначения, которые нам потребуются в дальнейшем:

- $H = \{\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{T-1}\}$ — последовательность наблюдений случайного процесса, где \mathbf{h}_t — вектор наблюдения, $\mathbf{h}_t \in R^n$, $\forall t$;
- S — множество скрытых состояний с элементами s_i ;
- N — количество элементов во множестве S ;

- q_t — текущее скрытое состояние, т.е. для любого момента времени t существует состояние $s_j \in S$ такое, что $q_t = s_j$;
- $g_i = p(q_t = s_i)$ — элементы вектора вероятностей G появления состояния s_i в первый момент времени.

СММ является вероятностной моделью совместной вероятности группы произвольных переменных $\{\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{T-1}, q_0, q_1, \dots, q_{T-1}\}$. В случае ее применения принимаются два допущения относительно случайных переменных:

1) вероятность появления скрытой переменной в момент времени t непосредственно зависит только от скрытой переменной в момент времени $t - 1$ (гипотеза для цепи Маркова):

$$P(q_t | \mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{T-1}, q_0, q_1, \dots, q_{T-1}) = p(q_t | q_{t-1}).$$

Элементы матрицы переходных вероятностей $A = \{a_{ij} = p(q_t = s_i | q_{t-1} = s_j)\}$ должны удовлетворять нормальным стохастическим условиям

$$p(q_t = s_i | q_{t-1} = s_j) \geq 0, \quad \sum_{i=1}^N p(q_t = s_i | q_{t-1} = s_j) = 1.$$

Эти условия равносильны утверждению, что из каждого состояния множества S в каждый момент времени с вероятностью единица происходит переход в какое-либо состояние этого же множества S .

2) как правило, допускается, что вероятность текущего наблюдения \mathbf{h}_t зависит от текущего скрытого состояния и не зависит от других переменных:

$$P(\mathbf{h}_t | \mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{T-1}, q_0, q_1, \dots, q_{T-1}) = p(\mathbf{h}_t | q_t).$$

Элементы вектора условных вероятностей $B = \{b_j(\mathbf{h}_t) = p(\mathbf{h}_t | q_t = s_j)\}$ должны удовлетворять условиям вида

$$\int_{R^n} p(\mathbf{h}_t | s_j) d\mathbf{h}_t = 1 \quad \forall j = 1 \dots N,$$

которые равносильны условию нормировки функции плотности вероятности.

СММ задана, если на множестве S задан полный набор ее параметров, т.е. $\lambda = \{G, B, A\}$.

В общем случае вероятности $p(q_t = s_i | q_{t-1} = s_j)$ и $p(\mathbf{h}_t | q_t = s_i)$ зависят от момента времени. Здесь не будет рассматриваться эта зависимость, ограничимся лишь однородным случаем.

Использование СММ предполагает решение трех основных проблем:

1. Расчет вероятности данной последовательности наблюдений для данной СММ $p(H | \lambda)$;
2. Поиск наилучшей последовательности состояний из множества S при заданной последовательности наблюдений H и параметрах модели λ ;
3. Поиск параметров модели λ при заданных последовательностях наблюдений H и скрытых состояний Q .

1.1. Алгоритм прямого-обратного хода

Расчет вероятности данной последовательности наблюдений для данной СММ $p(H | \lambda)$ относительно прост. Необходимо определить вероятность последовательности наблюдений H при некоторой фиксированной последовательности состояний $Q = \{q_0, \dots, q_{T-1}\}$ и просуммировать полученную вероятность по всевозможным последовательностям состояний. Считая, что все последовательности состояний равновероятны, искомую вероятность можно записать в виде

$$P(H | \lambda) = \sum_Q g_{q_0} b_{q_0}(\mathbf{h}_0) a_{q_0 q_1} b_{q_1}(\mathbf{h}_1) \dots a_{q_{T-2} q_{T-1}} b_{q_{T-1}}(\mathbf{h}_{T-1}). \quad (\text{П. 1.1})$$

Теоретически на формуле (П. 1.1) можно было бы остановиться, но практически количество операций, которое требуется для вычисления, не позволяет ее использовать. Для сокращения количества операций более пригоден итерационный метод вычислений, названный алгоритмом прямого-обратного хода.

Переменную $\alpha_t(i)$ назовем прямой переменной, и пусть $\alpha_t(i) = P(\mathbf{h}_0 \mathbf{h}_1 \dots \mathbf{h}_t, q_t = s_i | \lambda)$ есть вероятность появления для данной модели λ частичной последовательности наблюдений

$\mathbf{h}_0 \mathbf{h}_1 \dots \mathbf{h}_t$ (до момента t и состояния в этот момент). Таким образом, возможно вычислить $\alpha_t(i)$ (для всех возможных i):

1. Начальное значение $\alpha_0(i) = g_i b_i(\mathbf{h}_0)$, $i=1,2,\dots,N$.
2. Значения для последующих моментов времени

$$\alpha_{t+1}(j) = \left[\sum_{i=0}^{N-1} \alpha_t(i) a_{ij} \right] b_j(\mathbf{h}_{t+1}), \quad j = 1,2,\dots,N; \quad t = 1,2,\dots,T-1.$$

3. В заключение можем вычислить вероятность $P(H | \lambda)$:

$$P(H | \lambda) = \sum_{i=0}^{N-1} \alpha_T(i).$$

В силу симметрии можно аналогично ввести обратную переменную $\beta_t(i)$: $\beta_t(i) = P(\mathbf{h}_{t+1} \mathbf{h}_{t+2} \dots \mathbf{h}_{T-1} | q_t = s_i, \lambda)$, суть которой — это вероятность появления для данной модели λ частичной последовательности наблюдений $\mathbf{h}_{t+1} \mathbf{h}_{t+2} \dots \mathbf{h}_{T-1}$ (от момента $t+1$ до T), если состояние в момент t есть s_i .

Вычисляя $\beta_t(i)$, имеем:

1. Начальное значение $\beta_T(i) = 1$, на этом шаге вероятность определяется произвольно.
2. Значения для предыдущих моментов времени

$$\beta_t(i) = \sum_{j=0}^{N-1} a_{ij} b_j(\mathbf{h}_{t+1}) \beta_{t+1}(j), \quad j=1,2,\dots,N; \quad t=1,2,\dots,T-1,$$

где t пробегает последовательно все состояния $t = T-1, T-2, \dots, 1$

3. В заключение можем вычислить вероятность $P(H | \lambda)$:

$$P(H | \lambda) = \sum_{i=0}^N g_i b_i(o_i) \beta_1(i).$$

1.2. Алгоритм динамического программирования

Поиск максимума вероятности $P(H | Q, \lambda)$ при известной последовательности наблюдений H и параметрах модели λ заключается в том, чтобы среди всевозможных последовательностей состояний из множества S найти такую последовательность, которая бы с максимальной вероятностью описывала наблюдаемый процесс. Формально эту задачу можно определить в следу-

ющей форме: вероятность какой-либо последовательности скрытых состояний Q^* на интервале $[0; T-1]$, состоящую из номеров множества $S \{k_0, k_1, \dots, k_{T-1}\}$ можно записать в виде

$$P(H | Q^*, \lambda) = g_{k_0} b_{k_0}(\mathbf{h}_0) \prod_{t=1}^{T-1} b_{k_t}(\mathbf{h}_t) a_{k_t k_{t-1}},$$

необходимо найти такую последовательность Q , для которой справедливо

$$Q = \underset{\{k_0, k_1, \dots, k_{T-1}\}}{\operatorname{argmax}} g_{k_0} b_{k_0}(\mathbf{h}_0) \prod_{t=1}^{T-1} b_{k_t}(\mathbf{h}_t) a_{k_t k_{t-1}}. \quad (\text{П. 1.2})$$

Будем называть такую последовательность НВ-траектория. Для ее вычисления используется метод динамического программирования [17].

Логарифмируя числитель (П. 1.2) получим

$$L = \rho_0(k_0) + \sum_{t=1}^{T-1} \rho_t(k_t, k_{t-1}), \quad (\text{П. 1.3})$$

где

$$\begin{aligned} \rho_0(k_0) &= \ln g_{k_0} + \ln b_{k_0}(\mathbf{h}_0); \\ \rho_t(w_t, w_{t-1}) &= \ln b_{k_t}(\mathbf{h}_t) + \ln a_{k_t k_{t-1}}. \end{aligned}$$

Последовательно для каждого момента времени будем находить максимумы функционала (П. 1.3) в виде

$$\begin{aligned} \Phi_0(k_0) &= \rho_0(k_0) \\ \Phi_1(k_1) &= \max_s (\Phi_0(k_0) + \rho_1(k_1, k_0)) \\ &\dots \\ \Phi_{T-1}(k_{T-1}) &= \max_s (\Phi_{T-2}(k_{T-2}) + \rho_{T-1}(k_{T-1}, k_{T-2})). \end{aligned}$$

Практически это означает, что для каждого элемента множества состояний в каждый момент времени мы находим лишь один переход, который доставляет функционалу (П. 1.3) локальный максимум. Возникает вопрос, возможно ли в таком алгоритме на некотором этапе проигнорировать такую траекторию, которая впоследствии доставит функционалу (П. 1.3) глобальный максимум.

Рассмотрим схему, показанную на рисунке П. 1.1. Количество состояний в столбце равно количеству состояний во множестве S , а количество столбцов равно количеству тактов дискретного времени T .

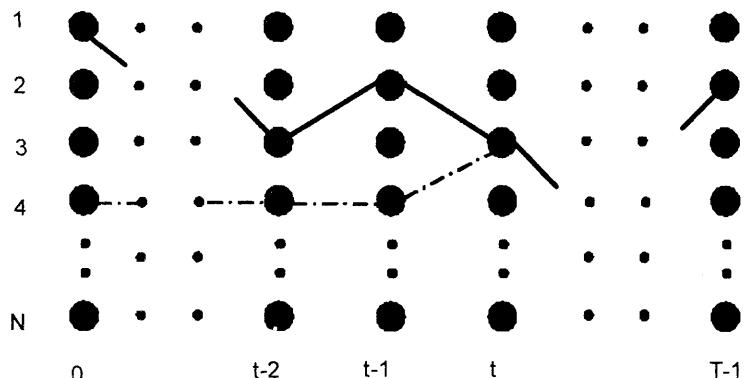


Рис. П.1.1. Схема возможных траекторий случайного процесса

Пусть нам известна НВ-траектория случайного процесса, и пусть величина значения максимума функционала (П. 1.3) на этой траектории $L_{\text{ист}, T}$. На рисунке П. 1.1 она выделена сплошной линией. Выберем некоторый момент времени t и предположим, что в этот момент НВ-траектория может быть отброшена и выбрана другая ложная траектория. На рисунке П. 1.1 она выделена штрих-пунктирной линией. Для этого необходимо, чтобы выполнялось условие $L_{\text{ист}, t} < L_{\text{лож}, t}$, где $L_{\text{ист}, t}$, $L_{\text{лож}, t}$ — значения функционала (П. 1.3) НВ- и ложной траектории в момент времени t , соответственно. На интервале $[t+1, T-1]$ эти траектории совпадают, и величина функционала (П. 1.3) на этом интервале равна p . Но тогда по условию максимума НВ-траектории и из свойства аддитивности функционала должно выполняться неравенство $L_{\text{ист}, t} + p > L_{\text{лож}, t} + p$, что приводит к неравенству $L_{\text{ист}, t} > L_{\text{лож}, t}$, которое противоречит нашему предположению. Это означает, что предложенный алгоритм всегда приводит к глобальному максимуму и

$$\max_Q L = \max_s \Phi_{T-1}(k_{T-1}),$$

где в левой части равенства максимум взят по всевозможным последовательностям состояний Q , а так же то, что последний элемент

$$k_{T-1}^* = \operatorname{argmax}_s \Phi_{T-1}(k_{T-1})$$

совпадает с последним элементом НВ-траектории.

Далее, используя рекуррентное правило,

$$\begin{aligned} m_t(k_t) &= \operatorname{argmax}_s (\Phi_{t-1}(k_{t-1}) + p_t(k_t, k_{t-1})), \\ k_t^* &= m_{t+1}(k_{t+1}^*) \end{aligned}$$

найдем всю последовательность состояний, в которой функционал (П. 1.3) имеет глобальный максимум.

1.3. Алгоритм Баума—Уолша

В СММ алгоритм Баума—Уолша используется для поиска параметров модели λ . Его идея основана на применении метода максимального правдоподобия.

Определим величину

$$\gamma_i(t) = p(q_t = s_i | H, \lambda),$$

которая означает вероятность того, что при известной последовательности наблюдений H скрытый процесс находится в состоянии s_i в момент времени t .

Если учесть положения, что

$$p(q_t = s_i | H, \lambda) = \frac{p(H, q_t = s_i | \lambda)}{P(H | \lambda)} = \frac{p(H, q_t = s_i | \lambda)}{\sum_{j=1}^N p(H, q_t = s_j | \lambda)},$$

и для переменных прямого и обратного хода справедливо

$\alpha_i(t)\beta_i(t) = p(\mathbf{h}_0, \dots, \mathbf{h}_t, q_t = s_i | \lambda)p(\mathbf{h}_{t+1}, \dots, \mathbf{h}_{T-1} | q_t = s_i, \lambda) = p(H, q_t = s_i | \lambda)$,
то в терминах этих переменных можно записать

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^N \alpha_j(t)\beta_j(t)}.$$

Введем величину вероятности, что при известной последовательности наблюдения скрытый процесс в момент времени t находится в состоянии s_i , а в момент времени $t+1$ в состоянии s_j

$$\varphi_{ij}(t) = \frac{p(q_t = s_i, q_{t+1} = s_j, H | \lambda)}{p(H | \lambda)} = \frac{\alpha_i(t)a_{ij}b_j(\mathbf{h}_{t+1})\beta_j(t+1)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_i(t)a_{ij}b_j(\mathbf{h}_{t+1})\beta_j(t+1)}.$$

Заметим, что введенная вероятность может быть выражено посредством величины $\gamma_i(t)$

$$\begin{aligned} \varphi_{ij}(t) &= \frac{p(q_t = s_i | H)p(\mathbf{h}_{t+1}, \dots, \mathbf{h}_{T-1}, q_{t+1} = s_j | q_t = s_i, \lambda)}{p(\mathbf{h}_{t+1}, \dots, \mathbf{h}_{T-1} | q_t = s_i, \lambda)} = \\ &= \frac{\gamma_i(t)a_{ij}b_j(\mathbf{h}_{t+1})\beta_j(t+1)}{\beta_i(t)}. \end{aligned}$$

Если просуммировать величину $\gamma_i(t)$ по времени, то мы получим относительное число моментов времени, в которые последовательность скрытых состояний находится в состоянии s_i . Сходным образом, если просуммировать величину $\varphi_{ij}(t)$ по времени, то получим относительное число переходов из состояния скрытой последовательности из состояния s_i в состояние s_j при известной последовательности H .

Используя эти значения относительных частот встреч скрытых состояний и переходов между ними, мы можем определить правила вычислений параметров СММ следующим образом:

$\tilde{g}_i = \gamma_i(0)$ — относительная частота встреч состояния s_i в момент времени 0;

$\tilde{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} \varphi_{ij}(t)}{\sum_{i=0}^{T-1} \gamma_i(t)}$ — ожидаемое число переходов из состояния s_i в состояние s_j относительно ожидаемого общего числа переходов из состояния s_i ;

$$\tilde{b}_i(k) = \frac{\sum_{t=0}^{T-1} \delta(\mathbf{h}_t = v_k) \gamma_i(t)}{\sum_{t=0}^{T-1} \gamma_i(t)} - \text{ожидаемое число раз, когда на-}$$

блюдение в состоянии s_i на выходе было равно v_k наблюдению, относительно общего числа наблюдений в этом состоянии (это выражения записано для дискретных распределений).

Алгоритм Баума—Уолша является рекуррентным и при его реализации очень важен шаг выбора начального приближения. Если нет каких-либо определенных соображений относительно начального приближения, то проще всего предполагать, что скрытые состояния и взаимные переходы между ними равновероятны, а наблюдаемая последовательность разбита на интервалы одинаковой длительности, каждый из которых соответствует определенному скрытому состоянию.

ПРИЛОЖЕНИЕ 2

EM-алгоритм

Определим общую задачу оценки параметров плотности распределения вероятности методом максимального правдоподобия. Пусть для описания наблюдаемого процесса $X = \{x_1, x_2, \dots, x_N\}$, $(x_i \in R^n, \forall i)$ была выбрана некоторая параметрическая функция плотности $p(X|\Theta)$, где Θ — множество параметров (соображения, на основании которых выбирается та или иная функция плотности, мы здесь обсуждать не будем). Для поиска параметров этой плотности распределения воспользуемся естественным предположением, что именно данная фиксированная наблюдаемая последовательность X является наиболее вероятной. При условии, что данные в последовательности X независимы и фиксированы, можно ввести функцию правдоподобия, зависящую от параметров Θ

$$P(X|\Theta) = \prod_{i=1}^N p(x_i|\Theta) = L(\Theta|X).$$

Максимизация функции L позволит найти параметры Θ^* , при которых этот максимум достигается

$$\Theta^* = \operatorname{argmax}_{\Theta} L(\Theta|X).$$

На практике чаще максимизируют логарифм функции правдоподобия $\log L(\Theta|X)$, поскольку это проще аналитически.

В зависимости от вида $p(x|\Theta)$ эта проблема может решаться легче или сложнее. Для примера, если $p(x|\Theta)$ — это просто одномерное нормальное распределение, где $\Theta = (\mu, \sigma)$, то эти значения возможно получить, дифференцированием $\log L(\mu, \sigma|X)$ по параметрам и решением системы из пары линейных уравнений относительно среднего и дисперсии. В большинстве случаев невозможно найти простое аналитическое выражение и необходимо прибегать к разработке более изощренных методов.

EM-алгоритм является именно таким методом. Этот алго-

ритм является общим методом нахождения оценки параметров, определяющих плотность распределения вероятностей с помощью максимума правдоподобия при условии недостатка предоставленных данных или их пропадания.

Существует два основных приложения ЕМ-алгоритма. Первое из них возникает в случае недостатка данных благодаря ограничениям в процессе наблюдения. Второе, когда оптимизация функции правдоподобия аналитически трудоемка, но сама функция правдоподобия может быть упрощена дополнительными предположениями относительно пропавших (или скрытых) параметров.

Предположим, что данные X наблюдаемы и сгенерированы некоторым распределением. Мы называем X неполными данными. Далее предположим, что существует полное множество данных $Z=(X, Y)$ и что полная функция плотности имеет вид

$$p(Z | \Theta) = p(X, Y | \Theta) = p(Y | X, \Theta)p(X | \Theta).$$

Такая форма представления функции плотности отражает отношения между скрытыми и наблюдаемыми данными.

С помощью этой новой функции плотности определим новую функцию правдоподобия $L(\Theta | Z) = L(\Theta | X, Y) = p(X, Y | \Theta)$, называемую правдоподобием полных данных. Заметим, что эта функция фактически случайная величина неизвестной скрытой информации Y , управляющая свойствами распределения. Предположим, что $L(\Theta | X, Y) = h_{X, \Theta}(Y)$, где $h_{X, \Theta}(Y)$ — некоторая функция случайной переменной Y при X и Θ константах. Исходная же функция правдоподобия $L(\Theta | X)$ представлена как функция правдоподобия неполных данных.

ЕМ-алгоритм на первом шаге (Е-шаг) находит ожидаемое значение логарифмического правдоподобия полных данных $\log p(X, Y | \Theta)$ при неизвестных данных Y и наблюдаемых данных X текущими параметрическими оценками.

Введем следующее выражение:

$$Q(\Theta, \Theta^{(i-1)}) = E[\log p(X, Y | \Theta) | X, \Theta^{(i-1)}], \quad (\text{П. 2.1})$$

где $\Theta^{(i-1)}$ — текущие параметрические оценки, которые исполь-

зуются при вычислении математического ожидания и Θ — новые параметры, которые оптимизируют Q .

Ключ к пониманию (П. 2.1) состоит в том, что X и $\Theta^{(i-1)}$ это константы, Θ — нормальная переменная, которую необходимо настроить. Правая часть уравнения (П. 2.1) может быть переписана в виде

$$E[\log p(X, Y | \Theta) | X, \Theta^{(i-1)}] = \int_{\mathcal{Y} \in \Theta} \log p(X, Y | \Theta) f(Y | X, \Theta^{(i-1)}) dY, \quad (\text{П. 2.2})$$

где $f(Y | X, \Theta^{(i-1)})$ — маргинальное распределение для ненаблюдаемых данных в зависимости как от наблюдаемых данных X , так и текущих значений параметров $\Theta^{(i-1)}$; \mathcal{Y} — пространство значений Y . В лучшем случае, маргинальное распределение имеет простую аналитическую зависимость от параметров $\Theta^{(i-1)}$ и возможных данных. В худшем — эта плотность очень сложна для анализа.

Второй шаг (M -шаг) ЕМ-алгоритма максимизирует выражение, рассчитанное на первом шаге, т.е. мы находим

$$\Theta^{(i)} = \underset{\Theta}{\operatorname{argmax}} Q(\Theta, \Theta^{(i-1)}).$$

Эти два шага принимаются как необходимые. Каждая итерация гарантирует рост логарифма правдоподобия и алгоритм гарантирует сходимость к локальному максимуму функции правдоподобия. Вопрос о сходимости здесь не обсуждается.

Задача оценки параметров смеси плотностей вероятностей — одно из наиболее широко используемых приложений ЕМ-алгоритма в распознавании речи. Пусть смесь предполагается заданной в виде

$$p(x | \Theta) = \sum_{i=1}^M \alpha_i p_i(x | \theta_i),$$

где вектор параметров $\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$ такой, что $\sum_{i=1}^M \alpha_i = 1$ и каждая p_i — параметрическая функция плотности.

Выражение логарифмического правдоподобия при неполных данных для этой формы плотности можно записать в виде

$$\log \prod_{i=1}^N p(x_i | \Theta) = \sum_{i=1}^N \log \left(\sum_{j=1}^M \alpha_j p_j(x_i | \theta_j) \right).$$

Это выражение трудно оптимизировать, так как оно содержит логарифм суммы. Если рассматривать данные X как неполные, но постулировать существование ненаблюдаемых событий $Y = \{y_i\}_{i=1}^N$, чьи значения сообщают, какой компоненте плотности соответствует каждое наблюдаемое значение, то выражение для правдоподобия значительно упрощается. Предположим, что $y_i \in \{1, \dots, M\}$ для каждого i и $y_i = k$, если i -й отчет был сгенерирован k -ой компонентой смеси. Если известны значения y , то выражение для правдоподобия приводится к виду

$$\log P(X, Y | \Theta) = \sum_{i=1}^N \log(P(x_i | y_i)P(y_i)) = \sum_{i=1}^N \log(\alpha_{y_i} p_{y_i}(x_i | \theta_{y_i})).$$

Данная частичная форма компонент плотности может быть оптимизирована с использованием вариационного метода.

Проблема состоит в том, что нам неизвестны значения y . Однако ее можно решить, если предположить, что y случайный вектор.

Далее необходимо получить выражение для распределения ненаблюдаемых значений. Выдвинем гипотезу относительно параметров смеси плотностей, т.е. что $\Theta^g = (\alpha_1^g, \dots, \alpha_M^g, \theta_1^g, \dots, \theta_M^g)$ — параметры, определяющие правдоподобие $L(\Theta^g | X, Y)$. При данных Θ^g можно легко вычислить $p_j(x_i | \theta_j^g)$ для каждого i и j . В дополнение смешивающие параметры α_j могут быть определены как априорные вероятности каждой компоненты смеси, т.е. $\alpha_j = p_j$. Используя правило Байеса, можно рассчитать

$$p(y_j | x_i, \Theta^g) = \frac{\alpha_j^g p_{y_j}(x_i | \theta_j^g)}{p(x_i | \Theta^g)} = \frac{\alpha_j^g p_{y_j}(x_i | \theta_j^g)}{\sum_{k=1}^M \alpha_k^g p_k(x_i | \theta_k^g)}$$

и

$$p(Y | X, \Theta^g) = \prod_{i=1}^N p(y_i | x_i, \Theta^g),$$

где элементы последовательности ненаблюдаемых данных $Y = \{y_1, \dots, y_N\}$ рассматриваются как независимые. Вернемся к уравнению (П. 2.2). Как видно, предположив существование скрытых переменных и построив гипотезу об исходных параметрах распределения, получена искомая граничная плотность.

Подставляя найденное маргинальное распределение в уравнение (П. 2.1) получим:

$$\begin{aligned} Q(\Theta, \Theta^g) &= \sum_{y \in Y} \log(L(\Theta | X, y)) p(y | X, \Theta^g) = \\ &= \sum_{y \in \varphi} \sum_{i=1}^N \log(\alpha_{y_i} p_{y_i}(x_i | \theta_{y_i})) \prod_{j=1}^N p(y_j | x_j, \Theta^g) = \\ &= \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \sum_{i=1}^N \log(\alpha_{y_i} p_{y_i}(x_i | \theta_{y_i})) \prod_{j=1}^N p(y_j | x_j, \Theta^g) = \\ &= \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \sum_{i=1}^N \delta_{l,y_i} \log(\alpha_l p_l(x_i | \theta_l)) \prod_{j=1}^N p(y_j | x_j, \Theta^g) = \\ &= \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l p_l(x_i | \theta_l)) \sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \delta_{l,y_i} \prod_{j=1}^N p(y_j | x_j, \Theta^g). \quad (\text{П. 2.3}) \end{aligned}$$

Это выражение для $Q(\Theta, \Theta^g)$ выглядит довольно громоздко, но его можно значительно упростить.

Во-первых, заметим, что для всех $l \in \{1, \dots, M\}$

$$\begin{aligned} &\sum_{y_1=1}^M \sum_{y_2=1}^M \dots \sum_{y_N=1}^M \delta_{l,y_i} \prod_{j=1}^N p(y_j | x_j, \Theta^g) = \\ &= \left(\sum_{y_1=1}^M \dots \sum_{y_{i-1}=1}^M \sum_{y_{i+1}=1}^M \dots \sum_{y_N=1}^M \prod_{j=1, j \neq i}^N p(y_j | x_j, \Theta^g) \right) p(l | x_i, \Theta^g) = \\ &= \prod_{j=1, j \neq i}^N \left(\sum_{y_j=1}^M p(y_j | x_j, \Theta^g) \right) p(l | x_i, \Theta^g) = p(l | x_i, \Theta^g), \quad (\text{П. 2.4}) \end{aligned}$$

так как $\sum_i p(i | x_i, \Theta^g) = 1$.

Воспользовавшись уравнением (П. 2.4), запишем уравнение (П. 2.3) следующим образом:

$$\begin{aligned} Q(\Theta, \Theta^g) &= \sum_{l=1}^M \sum_{i=1}^M \log(\alpha_l p_l(x_i | \theta_i)) p(l | x_i, \Theta^g) = \\ &= \sum_{l=1}^M \sum_{i=1}^M \log(\alpha_l) p(l | x_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log(p_l(x_i | \theta_i)) p(l | x_i, \Theta^g). \quad (\text{П. 2.5}) \end{aligned}$$

Чтобы максимизировать это выражение, можно отдельно максимизировать первое слагаемое, включающее только α_l , и второе слагаемое, включающее только θ_i , поскольку они не связаны.

Чтобы максимизировать выражение для α_l , введем множитель Лагранжа λ с ограничением $\sum_l \alpha_l = 1$, и вычислим следующее выражение:

$$\frac{\partial}{\partial \alpha_l} \left[\sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | x_i, \Theta^g) + \lambda \left(\sum_l \alpha_l - 1 \right) \right] = 0$$

или

$$\sum_{i=1}^N \frac{1}{\alpha_l} p(l | x_i, \Theta^g) + \lambda = 0.$$

Суммируя последнее выражение по l , получим $\lambda = -N$, что в результате дает

$$\alpha_l = \frac{1}{N} \sum_{i=1}^N p(l | x_i, \Theta^g).$$

Для некоторых распределений, можно получить аналитические выражения относительно θ_i . Например, для d -мерного распределения Гаусса с вектором средних значений μ и ковариационной матрицей Σ , $\theta = (\mu, \Sigma)$:

$$p_l(x | \mu_l, \Sigma_l) = \frac{1}{(2\pi)^{d/2} \det(\Sigma_l)} e^{-\frac{1}{2}(x - \mu_l)^T \Sigma_l^{-1} (x - \mu_l)}. \quad (\text{П. 2.6})$$

Чтобы вывести уравнения для параметров θ_i , вспомним некоторые положения алгебры матриц.

След квадратной матрицы $\text{tr}(A)$ равен сумме A диагональных элементов. След скаляра равен этому скаляру. А также $\text{tr}(A+B) = \text{tr}(A)+\text{tr}(B)$ и $\text{tr}(AB) = \text{tr}(AB)$, т.е. подразумевается, что $\sum_i x_i^T A x_i = \text{tr}(AB)$, где $B = \sum_i x_i x_i^T$.

Нам потребуется взять производные функции матрицы $f(A)$

по элементам матрицы. Определим производную $\frac{\partial f(A)}{\partial A}$ как мат-

рицу с элементами $\left[\frac{\partial f(A)}{\partial a_{ij}} \right]$, где a_{ij} — это элемент матрицы A .

Из определения следует, что, во-первых,

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x;$$

во-вторых, когда A является симметричной матрицей,

$$\frac{\partial \det A}{\partial a_{ij}} = \begin{cases} A_{ij} & \text{если } i=j \\ 2A_{ij} & \text{если } i \neq j \end{cases}$$

где A_{ij} — минор матрицы A .

Из вышеприведенного выражения видно, что

$$\frac{\partial \log(\det A)}{\partial A} = \begin{cases} A_{ij} / \det A & \text{если } i=j \\ 2A_{ij} / \det A & \text{если } i \neq j \end{cases} = 2A^{-1} - \text{diag}(A^{-1}),$$

Это следует из определения инвертированной матрицы.

В итоге получим

$$\frac{\partial \text{tr}(AB)}{\partial A} = B + B^T - \text{diag}(B).$$

Возьмем логарифм уравнения (П. 2.6), игнорируя постоянные (так как они исчезнут в процессе вычисления производных), и подставим его в правую часть уравнения (П. 2.5):

$$\sum_{l=1}^M \sum_{i=1}^N \log(p_l(x_i | \mu_l, \Sigma_l)) p(l | x_i, \Theta^g) = \\ = \sum_{l=1}^M \sum_{i=1}^N \left(-\frac{1}{2} \log(\det \Sigma_l) - \frac{1}{2} (x_i - \mu_l)^T \Sigma_l^{-1} (x_i - \mu_l) \right) p(l | x_i, \Theta^g). \quad (\text{П. 2.7})$$

Вычисления производной уравнения (П. 2.7) относительно μ_l приводят к системе уравнений:

$$\sum_{i=1}^N \Sigma_l^{-1} (x_i - \mu_l) p(l | x_i, \Theta^g) = 0,$$

решение которой может быть записано в виде

$$\mu_l = \frac{\sum_{i=1}^N x_i p(l | x_i, \Theta^g)}{\sum_{i=1}^N p(l | x_i, \Theta^g)}.$$

Чтобы найти Σ_l , необходимо написать уравнение (П. 2.7) в виде

$$\sum_{l=1}^M \left[\frac{1}{2} \log(\det \Sigma_l^{-1}) \sum_{i=1}^N p(l | x_i, \Theta^g) - \frac{1}{2} \sum_{i=1}^N p(l | x_i, \Theta^g) \text{tr}(\Sigma_l^{-1} (x_i - \mu_l)(x_i - \mu_l)^T) \right] = \\ = \sum_{l=1}^M \left[\frac{1}{2} \log(\det \Sigma_l^{-1}) \sum_{i=1}^N p(l | x_i, \Theta^g) - \frac{1}{2} \sum_{i=1}^N p(l | x_i, \Theta^g) \text{tr}(\Sigma_l^{-1} U_{li}) \right],$$

где $U_{li} = (x_i - \mu_l)(x_i - \mu_l)^T$.

Вычисляя производную относительно Σ_l^{-1} , получим

$$\frac{1}{2} \sum_{i=1}^N p(l | x_i, \Theta^g) (2\Sigma_l - \text{diag}(\Sigma_l)) - \frac{1}{2} \sum_{i=1}^N p(l | x_i, \Theta^g) (2U_{li} - \text{diag}(U_{li})) = \\ = \frac{1}{2} \sum_{i=1}^N p(l | x_i, \Theta^g) (2M_{li} - \text{diag}(M_{li})) = 2S - \text{diag}(S),$$

где $M_{li} = \Sigma_l - U_{li}$; $S = \frac{1}{2} \sum_{i=1}^N p(l | x_i, \Theta^g) M_{li}$. Приравняв производную к нулю, т.е. $2S - \text{diag}(S) = 0$, получим $S = 0$. Это в свою очередь аналогично выражению

$$\sum_{i=1}^N p(l | x_i, \Theta^g) (\Sigma_l - U_{li}) = 0$$

или

$$\Sigma_l = \frac{\sum_{i=1}^N p(l | x_i, \Theta^g) U_{li}}{\sum_{i=1}^N p(l | x_i, \Theta^g)} = \frac{\sum_{i=1}^N p(l | x_i, \Theta^g) (x_i - \mu_l)(x_i - \mu_l)^T}{\sum_{i=1}^N p(l | x_i, \Theta^g)}.$$

В итоге для расчета новых параметров, функционально зависящих от старых, справедливы следующие формулы:

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^N p(l | x_i, \Theta^g);$$

$$\mu_l^{new} = \frac{\sum_{i=1}^N x_i p(l | x_i, \Theta^g)}{\sum_{i=1}^N p(l | x_i, \Theta^g)};$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^N p(l | x_i, \Theta^g) (x_i - \mu_l^{new})(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum_{i=1}^N p(l | x_i, \Theta^g)}.$$

ПРИЛОЖЕНИЕ 3

1. Метод выделения формант

На сегменте длительностью T последовательности данных $\{q_1, q_2, \dots, q_t\}$ с помощью ДПФ получим спектр, который сгладим рекурсивным фильтром Баттервортса 5-го порядка [33] с частотой среза 360 Гц. Нелинейный фазовый сдвиг устраним методом двухсторонней фильтрации для получения сглаженного спектра с гармониками вида h_m , где t — номер сегмента длительностью T , на котором получен спектр, n — номер гармоники спектра.

Опустим временной индекс у гармоник сглаженного спектра h_m и будем их в дальнейшем обозначать как h_n . Допустим, что гармоники можно аппроксимировать в виде

$$h_n \approx \sum_{j=1}^N Q_{nj}, \quad (\text{П. 3.1})$$

где функции определяются следующим образом:

$$Q_{nj} = \begin{cases} A_j e^{-\frac{(n-\omega_j)^2(\Delta\omega_j)^2}{\sigma_{j,1}}}, & n \leq m_j, \\ A_j e^{-\frac{(n-\omega_j)^2(\Delta\omega_j)^2}{\sigma_{j,2}}}, & n > m_j, \end{cases} \quad (\text{П. 3.2})$$

здесь индекс j нумерует шаг итерационных вычислений.

На первом шаге итерации

$$\left. \begin{array}{l} D_n^{(1)} = h_n \\ \omega_1 = \arg \max_n \{D_n^{(1)}\} \\ A_1 = \max_n \{D_n^{(1)}\} \end{array} \right\}, \quad (\text{П. 3.3})$$

параметры $\sigma_{j,1}$ и $\sigma_{j,2}$ вычисляются из условий:

$$\left. \begin{array}{l} Q_{n-g,1} = D_{n-g}^{(j)}, \\ Q_{n+g,1} = D_{n+g}^{(j)}, \end{array} \right\}, \quad (\text{П. 3.4})$$

где величина g выбирается из условия, что величина формантной области не менее 200 Гц, т.е. $g = 200/\Delta\omega$ и соответственно на j -ом шаге итерации

$$\left. \begin{array}{l} D_n^{(j)} = D_n^{(j-1)} - Q_{n,j-1} \\ \omega_j = \arg \max_n \{D_n^{(j)}\} \\ A_j = \max_n \{D_n^{(j)}\} \end{array} \right\}, \quad (\text{П. 3.5})$$

параметры $\sigma_{j,1}$ и $\sigma_{j,2}$ вычисляются из условий:

$$\left. \begin{array}{l} Q_{n-g,j} = D_{n-g}^{(j)} \\ Q_{n+g,j} = D_{n+g}^{(j)} \end{array} \right\}. \quad (\text{П. 3.6})$$

Итерационный процесс продолжается, пока не выполнится условие

$$(D_n^{(j-1)} - Q_{n,j-1})^2 < \varepsilon, \quad (\text{П. 3.7})$$

где ε — установленная точность аппроксимации сглаженного спектра.

На рисунках П. 3.1 — П. 3.4 представлены примеры сглаженных спектров (сплошная кривая) для гласных звуков «а», «и», «о» и «э», соответственно. Пунктирной кривой показаны гаусссоиды, с помощью которых аппроксимируется сглаженный спектр.

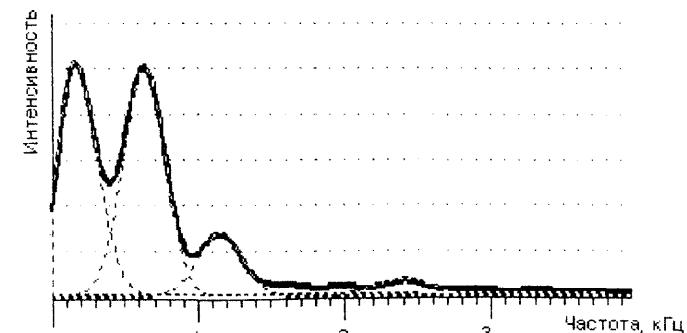


Рис. П. 3.1. Аппроксимация сглаженного спектра фонемы «а»

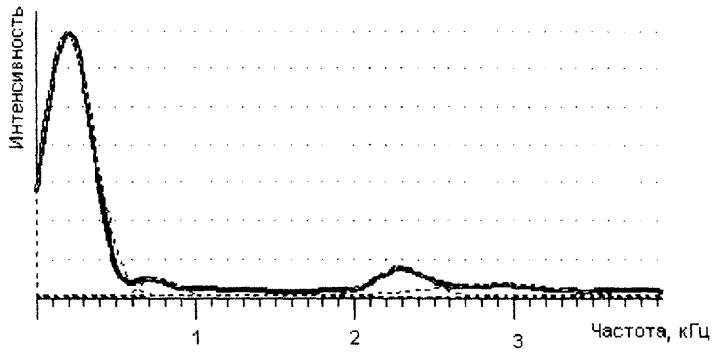


Рис. П. 3.2. Апроксимация сглаженного спектра фонемы «и»

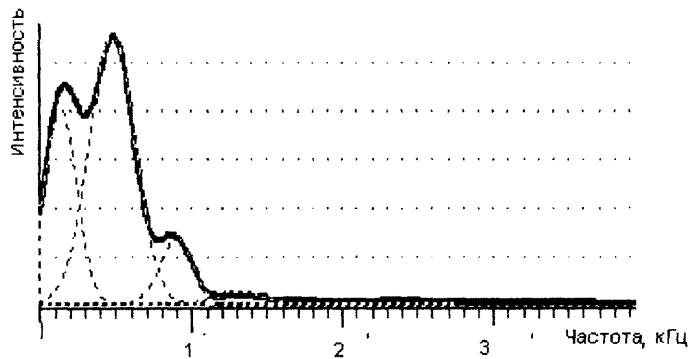


Рис. П. 3.3. Апроксимация сглаженного спектра фонемы «о»

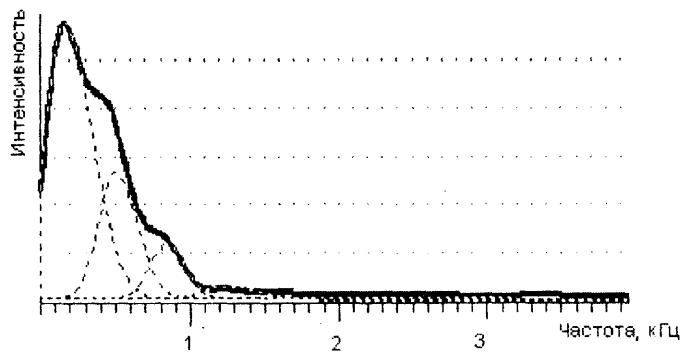


Рис. П. 3.4. Апроксимация сглаженного спектра фонемы «э»

Найденные в результате аппроксимации (П. 3.1) — (П. 3.6) значения (A, ω) будем называть амплитудами и частотами формант, а множество значений $\{(A, \omega)_k\}$, — формантным набором, где k — количество формант.

2. Метод сравнения формантных наборов

Проведем сравнение двух формантных наборов $L_i = \{(A, \omega)_k\}$ и $L_{i'} = \{(A, \omega)_k\}_{i'}$. Введем вероятностную меру сходства для пары параметров на множествах $\{(A, \omega)_k\}_i$ и $\{(A, \omega)_k\}_{i'}$:

$$P_{ij} = \frac{1}{2\pi\sigma_A\sigma_\omega} \exp\left\{-\frac{(A_j^{(i)} - A_i^{(i)})^2}{2\sigma_A^2} - \frac{(\omega_j^{(i)} - \omega_i^{(i)})^2}{2\sigma_\omega^2}\right\}, \quad (\text{П. 3.8})$$

где A и ω — независимые случайные величины, подчиняющиеся нормальному распределению; σ_A , σ_ω — дисперсии соответствующих величин (параметры модели); i, j порядковые номера параметров в соответствующих множествах L_i и $L_{i'}$.

Далее будем считать, что всякая форманта (A, ω) , наблюдавшаяся во множестве, может либо появится, либо исчезнуть, либо иметь свое продолжение в последующем и предыдущем наборах. Для того чтобы выявить поведение данной пары параметров, необходимо провести прореживание матрицы P .

Прореживание матрицы P состоит в том, что среди ее элементов находится максимальный, пусть он находится на пересечении i -ой строки и j -го столбца. Все элементы этого столбца и строки, кроме максимального, заменим нулями, в результате получим новую матрицу P' . Проведем эту же последовательность операций с новой матрицей, не рассматривая строку, в которой уже найден максимальный элемент, — получим следующую матрицу. Операции будем продолжать до тех пор, пока не исчерпаем все строки или столбцы, на каждом шаге запоминая значение максимального элемента и его расположение в матрице. Если значение максимального элемента меньше, чем заранее определенный порог, то считается, что соответствующий элемент матрицы равен нулю. Если матрица не квадратная,

то останутся лишние строки (столбцы). Фактически это означает, что количество формант во множестве L , больше (меньше) количества формант во множестве L_d . Формантам, которые исчезли или появились, в прореженной матрице будет соответствовать строка (столбец), состоящая из нулей.

Вероятность исчезновения пика или появления ранее не существовавшего пика определим как

$$P_i(A, \omega) = \frac{1}{\sigma_A \sqrt{2\pi}} \exp\left\{-\frac{A_i^2}{2\sigma_A^2}\right\}.$$

В качестве меры сходства двух формантных наборов используем величину

$$\xi(L_t, L_d) = \frac{\sum_{\Omega} P_{ij} (A_i^{(d)} + A_j^{(t)}) + \sum_{\Omega_1} P_i A_i^{(t)} + \sum_{\Omega_2} P_i A_i^{(d)}}{\sum_{i=0}^d A_i^{(d)} + \sum_{i=0}^t A_i^{(t)}}, \quad (\text{П. 3.9})$$

где Ω — множество формант, для которых элементы прореженной матрицы не равны нулю; Ω_1 , Ω_2 — множество возникших или исчезнувших формант во множествах L_t и L_d , соответственно.

На основе понятия прореженной матрицы можно ввести понятие формантной траектории, которое часто используется в данной работе. Если в последовательности формантных наборов L_t , L_{t+1} , ..., L_{t+n} можно найти такую последовательность формант $\{(A, \omega)_t, (A, \omega)_{t+1}, \dots, (A, \omega)_{t+n}\}$ (по одной паре на каждый набор), таких что для любых соседних пар из этой последовательности существует не нулевой элемент прореженной матрицы, то последовательность формант будем называть траекторией.

ЛИТЕРАТУРА

1. Шелухин О.И., Лукьянцев Н.Ф. Цифровая обработка и передача речи. М.: Радио и связь, 2000.
2. Agranovsky A.V., Lednov D.A. Repalov S.A. Working out voice warping compensation parameters estimation technique // IAFP&ENFSI International Association for Forensic Phonetics & European Network of Forensic Science Institutions (Speech and Audio Working Group) // Abstracts, Annual Conference, 1–4 July, 2002. P. 2–9.
3. Agranovsky A.V. Lednov D.A. Potapenko A.M. Repalov S.A. Segmentation of Signal, Containing a Talk of Several Speakers, into Monologue Components // International Workshop "SPEECH AND COMPUTER" (SPECOM'2001). P. 139–142.
4. Computational Auditory Scene Analysis / Eds. by D. Rosenthal and H. Okuno. Mahwah. N.Y.: Lawrence Erlbaum Associates, 1997.
5. Сердюков В.Д. Опознавание речевых сигналов на фоне мешающих факторов. Тбилиси: «Мецниереба», 1987.
6. Рамишвили Г.С. Автоматическое опознавание говорящего по голосу. М.: Радио и связь, 1981.
7. Фролов М.В. Контроль функционального состояния человека оператора. М.: Наука, 1985.
8. Jelinek F. Continuous speech recognition by statistical methods // Proc. IEEE. Apr. 1976. V. 64. P. 532–556.
9. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев: Наукова думка, 1987.
10. Аграновский А.В., Можаев О.Г., Леднов Д.А., Зулкарнеев М.Ю. Метод идентификации языка, основанный на фонетическом содержании сообщения // Мат-лы науч.-технич. конф., 16–20 сентября 2002, п. Кацивели, Крым, Украина. Таганрог: Изд-во ТРТУ, 2002. Т. 2. С. 29–32.
11. Большаков И.А. Статистические проблемы выделения потока сигналов из шума. М.: Советское радио, 1969.
12. Харкевич А.А. Борьба с помехами. 2-е изд. М.: Наука, 1965.
13. Вахитов Я.Ш. Слух и речь: Конспект лекций по курсу «Электроакустика». Раздел 2. Л., 1973.

14. Сорокин В.Н. Теория речеобразования. М.: Наука, 1985.
15. Narada Dilp Warakagoda A Hybrid ANN-HMM ASR system with NN based adaptive preprocessing // M.Sc. Thesis, Norges Tekniske Hogskole, Institutt for Teleteknikk, Transmisionsteknikk, May 19, 1996.
16. Bourlard H., Morgan N. Continuous speech recognition by connectionist statistical methods // IEEE Trans. On neural networks. Nov. 1993. V. 4. № 6. P. 893—909.
17. Момтья В.В., Мучник И.Б. Скрытые Марковские модели в структурном анализе сигналов. М.: ФИЗМАТЛИТ, 1999.
18. Сосулин Ю.Г. Теория обнаружения и оценивания стохастических сигналов. М.: Советское радио, 1978.
19. Коваленко И.Н., Кузнецов Н.Ю., Шуренков В.М. Случайные процессы: Справочник. Киев: Наукова думка, 1983.
20. Павловский З. Введение в математическую статистику. М.: Статистика. 1965.
21. Bilmes J.A. A Gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models // ICSI-TR-97-021, April 1998.
22. Аграновский А.В., Леднов Д.А. Рекурсивный алгоритм определения параметров нормального процесса авторегрессии для скрытых моделей Маркова в системах обработки речевых данных: Тез. докл. Четвертого Всероссийского Симпозиума по прикладной и промышленной математике // Обозрение прикладной и промышленной математики. Петрозаводск, 2003. Т. 10. Вып. 1. С. 84—86.
23. Ивахненко А.Г. и др. Принятие решений на основе самоорганизации. М.: Советское радио, 1976.
24. Page E.S. Continous inspection schemes // Biometrika. 1954. V. 41. P. 100—114.
25. Патрик Э. Основы теории распознавания образов // Пер. с англ. / Под ред. Б.Р. Левина. М.: Советское радио, 1980.
26. Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р. и др. Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989.
27. Computational Auditory Scene Analysis. Eds. by D. Rosenthal, H. Okuno. N.Y.: Lawrence Erlbaum Associates, 1997.
28. Slaney M., Naar D., Lyon R.F. Auditory Model Inversion For Sound Separation // Proceedings of the ICASSP 94 1994 International Conference on Acoustics, Speech, and Signal Processing. Adelaide, Australia, 19—22 April 1994.
29. Ottaviani L., Rocchesso D. Separation of Speech Signal From Complex Auditory Scenes // Proc. of the COST G-6 Conf. on Digital Audio Effects (DAFX-01). Limerick, Ireland, December 6—8, 2001.
30. Аграновский А.В., Куликов Л.С., Леднов Д.А. Цифровая обработка сигналов и формирование вектора признаков в задачах классификации речи. Ростов н/Д: Изд. СКНЦ ВШ, 2004.
31. Ландау Л.Д., Либфриц Е.М. Теоретическая физика. Т. VI. Гидродинамика. М.: ФИЗМАТЛИТ, 2003.
32. Вокодерная телефония. Методы и проблемы / Под ред. А.А. Пирогова. М.: Связь, 1974.
33. Gold B. Computer program for Pitch Extraction // JASA. 1962. V. 32. № 7. P. 916—921.
34. Manley H.J. Analysis-Synthesis of connected Speech in Terms of orthogonalised Exponentially Damped Sinusoid // JASA. 1963. № 4. V. 35. P. 464—474.
35. Хемминг Р.В. Цифровые фильтры // Пер. с англ.: Под ред. А.М. Трахтмана. М.: Советское радио, 1980.
36. Архипов И.О., Гитлин В.Б. Оценка точности выделения основного тона методом GS // XI сессия Рос. Ак. Общ. «Современные речевые технологии»: сб. трудов. М., 26—28 января 1999. С. 38—42.
37. Huici M.E.H.D., Ginori J.V.L. Combined algorithm for pitch detection of speech signals// Electronics Letters. 5th January 1995. V. 31. №. 1.
38. Рабинер Л., Голд Б. Теория и применение цифровой обработки сигналов. М.: Мир, 1978.
39. Аграновский А.В., Леднов Д.А., Потапенко А.М., Репалов С.А., Сулима П.М. Способ выделения основного тона из речевого сигнала // Пат. № 2184399 РФ, выд. 27.08.2002, приор. от 22.09.2000 по заяв. № 2000124181/09.

40. Аграновский А.В., Арутюнян Р.Э., Репалов С.А. Выделение монологических составляющих беседы многих дикторов// Сб. тр. XIII сессии РАО, 27-29 августа 2003. Т. 3. Акустика речи. Медицинская и биологическая акустика. М.: ГЕОС, 2003. С. 21—25.
41. Златоусова Л.В., Крячи С.А. Классификация мужских и женских голосов по акустическим характеристикам // Там же. С. 45—48.
42. Doddington G.R., Flanagan G.L., Lummis R.C. Automatic speaker verification by non-linear time alignment of acoustic parameters. Патент США № 3700815. 1972.
43. Doddington G.R., Hidrick B. Some results on speaker verification using amplitude spectra // JASA. 1974. V. 55. № 2.
44. Иванов А.И. Биометрическая идентификация личности по динамике подсознательных движений. Пенза, 2000.
45. Ming-Shih Chen, Pie-Hwa Lin, Hsiao-Chuan Wang. Speaker Identification Based on a Matrix Quantization Method // IEEE Trans. On Signal Proc. Jan. 1993. V. 41. № 1.
46. Furui S. Cepstral analysis technique for automatic speaker verification // IEEE Trans. Acoust., Speech, Signal Process. Apr. 1987. V. ASSP-29. P. 254—272.
47. Adami A.G., Hermansky H. Segmentation of Speech for Speaker and Language Recognition, // Proc. EUROSPEECH. 2003.
48. Jin Q., Schultz T., Waibel. Phonetic Speaker Identification // Proc. ICSLP. 2002.
49. Park A., Hazen T. J. ASR Depended Techniques for speaker identification // Proc. ICSLP. 2002.
50. He J., Liu L., Palm G. A new codebook training algorithm for VQ-based speaker recognition // IEEE Proc. of International Conference on Acoustics, Speech and Signal Processing. Munich, 1997. V. 2. P. 1091—1094.
51. Kohonen T. The self-organizing map // Proc of IEEE. V. 78. P. 1464—1480.
52. Дуда Р., Харт П. Распознавание образов и анализ сцен / Пер. с англ. Г.Г. Вайнштейна и А.М. Васьковского. Под ред. В.Л. Стефанюка. М.: Мир, 1976.
53. Doddington G.R., Przybocki M.A., Martin A.F., Reynolds D.A. The NIST speaker recognition evaluation: Overview, methodology, systems, results, perspective // Speech Communication. 2001. V. 31. P. 225—254.
54. Аграновский А.В., Леднов Д.А., Репалов С.А. Оценка точности текстонезависимых систем идентификации дикторов, на основе экспериментальных АЧХ голосовых трактов дикторов // Телекоммуникации. 2000. № 6. С. 3—17.
55. Аграновский А.В., Леднов Д.А. Репалов С.А. Метод текстонезависимой идентификации диктора на основе индивидуальности произношения гласных звуков // Акустика и прикладная лингвистика: Ежегодник РАО. Вып. 3. М., 2002. С. 103—115.
56. Oppenheim A. V., Schafer R. W. Discrete-Time Signal Processing. Prentice Hall, Englewood Cliffs. N.Y., 1989.
57. Atal B.S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification // Journal of the Acoustical Society of America, 1974. V. 55. P. 1304—1312.
58. Потапова Р.К. Речь: коммуникация, информация, кибернетика: Учебное пособие. М.: Эдиториал УРСС, 2001.
59. Кодзасов С.В., Кривнова О.Ф. Общая фонетика: Учебник. М.: Рос. гос. гуманит. ун-т, 2001.
60. Hwang M.-Y., Huang X., Alleva F. Predicting Unseen Triphones with Senones // Proc. ICASSP 1993. II. Minneapolis, USA P. 311—314.
61. Juang B.H., Levinson S., Sondhi M.M. Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains // IEEE Transactions on Information Theory. March 1986. V. IT-32. № 2. P. 307—309.
62. Chin K.K., Woodland P.C. Maximum Mutual Information Training of Hidden Markov Models with Vector Linear Predictors // Proc. ICSPL. 2002.
63. Винюк Т.К. Сравнение ИКДП- и НММ-методов распознавания речи // Методы и средства информ. речи. Киев, 1991.
64. Bellegarda J.R., Nahamoo D. Tied Mixture Continuous Parameter Modeling for Speech Recognition // IEEE Trans ASSP, 1990. V. 38. № 12. P. 2033—2045.

65. Huang X.D., Jack M.A. Semicontinuous hidden Markov models for Speech Signals // Computer Speech and Language. 1989. V. 3. № 3. P. 239—252.

66. Huang X.D., Hon H.W., Hwang M.Y., Lee K.F. A Comparative Study of Discrete, Semicontinuous and Continuous Hidden Markov Models // Computer Speech and Language. 1993. V. 7. № 4. P. 359—368.

67. Young S.J. The General Use of Tying in Phoneme-Based HMM Speech Recognisers // Proc ICASSP, San Francisco, March 1992. V. 1. P. 569—572.

68. Hwang M.Y., Huang X. Shared Distribution Hidden Markov Models for Speech Recognition // IEEE Trans Speech and Audio Processing. 1993. V. 1. № 4. P. 414—420.

69. Young S.J., Woodland P.C. State Clustering in HMM-based Continuous Speech Recognition // Computer Speech and Language. 1994. V. 8. № 4. P. 369—384.

70. Digalakis V., Monaco P., Murveit H. Genones: Generalised Mixture Tying in Continuous Speech HMM-based Speech Recognisers // IEEE Trans Speech and Audio Processing. 1995.

71. Bahl L.R., Souza P.V. de, Gopalakrishnan P.S., Nahamoo D., Picheny M.A. Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees // Proc. DARPA Speech and Natural Language Processing Workshop. Calif. Feb, 1991. P. 264—270.

72. Kannan A., Ostendorf M., Rohlicek J.R. Maximum Likelihood Clustering of Gaussians for Speech Recognition // IEEE Trans. on Speech and Audio Processing. 1994. V. 2. № 3. P. 453—455.

Учебное издание

Александр Владимирович АГРАНОВСКИЙ
Дмитрий Анатольевич ЛЕДНОВ

ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ АЛГОРИТМОВ
ОБРАБОТКИ И КЛАССИФИКАЦИИ
РЕЧЕВЫХ СИГНАЛОВ