Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

Кафедра электронно-вычислительных средств

Речевые интерфейсы ЭВС

Учебно-методическое пособие
для студентов специальности I – 40 02 02
«Электронные вычислительные средства»
дневной формы обучения

Минск 2005

УДК 004.5(075.8)
ББК 32.973 я 73
Р 30

Р е ц е н з е н т ы:
заведующий кафедрой ИИТ БГУИР, д-р техн. наук, проф. В.В. Голенков,
заведующий кафедрой ЭВМ БГУИР, д-р техн. наук, проф. Р.Х. Садыхов

А в т о р ы:
А.А. Петровский, Д.С. Лихачёв, Ал.А. Петровский,
А.В. Шадевский, М.З. Лившиц, А.Н. Павловец

Р 30 **Речевые** интерфейсы ЭВС: Учебно-метод. пособие для студ. спец. I – 40 02 02 «Электронные вычислительные средства» дневной формы обуч. / А.А. Петровский, Д.С. Лихачёв, Ал.А. Петровский и др. – Мн.: БГУИР, 2005. – 51 с.: ил.

ISBN 985-444-872-X

Учебно-методическое пособие содержит описание алгоритмов, применяемых для обработки речи: детектора речи, анализа на основе линейного предсказания, векторного квантования. Даны примеры применения векторного квантования для кодирования параметров в широкополосном вокодере на основе CELP-модели с многополосным возбуждением, перцептуальной оптимизацией и синусоидальном кодере речи на основе антропоморфического анализа.

Также приводятся указания по выполнению лабораторных работ.

УДК 004.5(075.8)
ББК 32.973 я 73

ISBN 985-444-872-X

Содержание

1. ДЕТЕКТОР РЕЧИ

1.1. Область применения и принцип действия

Детектор речи является важной частью современных приложений по обработке речи. Детектирование речи используется в системах кодирования и распознавания речи, а также в системах повышения ее качества. Характеристики детектора речи во многом определяют качество работы всей системы в целом. Поэтому алгоритмы детектирования часто являются наиболее критической частью таких систем и одновременно с улучшением их качества увеличиваются требования к алгоритмам детектирования речи.

Основной целью детектора речи является обеспечение эффективного определения наличия речи в поступающем на его вход сигнале на фоне изменяющейся акустической обстановки. В большинстве алгоритмов решение речь/пауза принимается на основе сравнения классификационного параметра с порогом. Одной из главных проблем является точное определение порога.

Когда уровень окружающего шума постоянен и отношение сигнал/шум (*SNR*) постоянно, проблема детектирования речи не выглядит такой уж сложной. Однако большинство систем телекоммуникации работают в присутствии окружающего шума. Встречаются нестационарные шумы с низким уровнем *SNR*. Это является большой проблемой в задачах детектирования речи. Первые алгоритмы детектирования требовали постоянного уровня окружающего шума. Они допускали большое количество ошибок и могли быть использованы только в простых приложениях. Поэтому следующим этапом в разработке алгоритмов детектирования речи стало создание адаптивных алгоритмов. Эти алгоритмы были более надежными в условиях нестационарных окружающих шумов. В детекторах на их основе характеристики шума определяются только в паузах.

Большинство алгоритмов генерируют ограниченную ошибку на выходе (короткий период молчания в речи и наоборот). Некоторые из этих ошибок связаны с малой энергией невокализованной речи, но их легко удалить из выходного потока данных. В современных детекторах речи для улучшения их характеристик используются двухшаговые алгоритмы или дополнительная постобработка выходной последовательности. Одним из таких методов является размытие выходной последовательности, целью которой является соединение коротких пауз между речевыми участками. Это также позволяет уменьшить вероятность ошибки на невокализованных участках. Однако такой механизм неэффективен при коррекции изолированных ошибок, особенно при больших интервалах тишины. Иногда используется медианная фильтрация, чтобы сгладить классификационный параметр. Этот метод подразумевает большую задержку и не избавляет от появления пиков в выходной последовательности.

1.2. Схемы реализации детектора речи

Процесс детектирования речи может быть представлен следующим образом: 1) поступающий на вход сигнал разделяется на сегменты; 2) для текущего сегмента рассчитывается вектор, определяющий значение выбранного классификационного параметра; 3) в зависимости от алгоритма определяется разница между значениями классификационного параметра для текущего и предыдущего фреймов, либо разница между значениями классификационного параметра и порога.

Порог вычисляется на основании простой статистики:

$$d_{thr} = mean(d) + l \cdot std(d), \qquad (1.1)$$

где d – определяет классификационный параметр, l – управляет доверительной границей. Среднее значение и стандартное отклонение определяется с помощью экспоненциального усреднения в паузах между речью. Текущий фрейм принимается активным, если значение классификационного параметра больше, чем порог.

В качестве классификационного параметра могут быть использованы различные параметры (энергия сигнала в фрейме, спектр, кепстр сигнала).

Прекращение оценки шума в речевых периодах является главным недостатком стандартных алгоритмов. Быстрое изменение уровня на речевых участках сигнала может привести к неверной работе детектора. Данный алгоритм работает только на ограниченном интервале стационарности шума. Поэтому для определения порога часто используется метод, называемый *минимум статистики*. Этот метод основан на отслеживании минимума классификационного параметра. Оценка классификационного параметра шума получается путем выбора минимального значения из L последних значений, где L соответствует интервалу стационарности шума.

Энергетический критерий оценки. Детектор речи, основанный на энергетической оценке сигнала, является наиболее простым. Энергия вычисляется для каждого фрейма сигнала по следующей формуле:

$$E = \sum_{n=0}^{N-1} | x(n) |^2 , \qquad (1.2)$$

где N – длина фрейма и $x(n)$ – входной речевой сигнал. В данном случае классификационным параметром является энергия сигнала. Порог вычисляется только в паузах согласно формуле (1.1).

Заметим, что среднее значение и стандартное отклонение, используемое при вычислении порога, определяется характеристиками окружающего шума.

Спектральная оценка. В качестве классификационного параметра используется оценка кратковременной спектральной плотности мощности в каждом фрейме, которая может быть получена с помощью дискретного преобразования Фурье (*ДПФ*).

Кепстральная оценка. Данный тип детекторов используют спектраль-

ные коэффициенты в качестве классификационных параметров. Быстроменяющиеся коэффициенты спектрального вектора содержат информацию об огибающей амплитуды речевого спектра. Кепстральный вектор может быть легко получен с помощью обратного *ДПФ* от логарифма амплитуды спектра, полученного с помощью прямого *ДПФ*. Однако характеристики данного метода часто ухудшаются при работе с сигналами с *SNR* около 0 дБ.

1.3. Применение принципов антропоморфизма при детектировании речи

В настоящее время одним из быстро развивающимся направлением в инженерии является нейроморфная инженерия, сущность которой заключается в разработке искусственных нейронных систем, чьи физическая архитектура и принципы конструирования взяты у биологических прототипов. Нейроморфная инженерия применяет принципы, взятые у биологических организмов для выполнения задач, с которыми эти организмы удачно справляются, но которые тяжело решить традиционными инженерными средствами.

В последнее время алгоритмы анализа речевого сигнала, основанные на свойствах человеческого уха, получили широкое распространение. Разработка такого класса методов базируется на предположении, что анализ речевого сигнала, основанный на слуховой модели человека, будет более успешный, чем анализ, основанный на абстрактных моделях восприятия или статистических марковских моделях. Условно процесс восприятия человеком речи можно разделить на два этапа: начальный и основной. На начальном этапе осуществляется преобразование акустического сигнала во внутреннее нейронное представление, в основе которого лежит слуховая спектрограмма. На основном этапе происходит анализ спектрограммы, в результате которого из нее извлекается контекст спектральных и временных модуляций.

На начальном этапе звуковой сигнал попадает в ухо и с помощью барабанной перепонки и косточек среднего уха преобразуется в механические колебания. Эти колебания возбуждают сложные пространственно-временные колебания вдоль основной мембраны слуховой улитки. Характер этих колебаний таков, что разные звуковые частоты преобразуются в активность, локализованную в различных точках звуковой мембраны. Таким образом, основная мембрана может быть представлена как банк фильтров с высокой степенью перекрытия равномерно распределенных на логарифмической частотной оси.

На основном этапе анализа слуховой спектр преобразуется в более подходящую форму, интерпретируется и разделяется на различные компоненты и параметры, связанные с различными источниками сигнала. На этой стадии производится оценка контекста слуховой спектрограммы. В этом процессе большую роль играют низкочастотные модуляционные составляющие, которые являются основными носителями информации в речевом сигнале и тембра в музыке. Другими словами, речь и другие аудиосигналы в действительности являются низкочастотными процессами, которые модулируются несущими частотами. Более 95 % модуляционных компонент речевого сигнала, влияющих на

разборчивость и восприятие речи, находятся в диапазоне от 1 до 16 Гц, с пиком около 3-5 Гц. Это обусловлено количеством слогов, произносимых человеком за одну секунду. Таким образом, модуляционные компоненты (шумы, реверберация), изменяющиеся с частотами, не входящими в данный диапазон, могут быть удалены с помощью фильтрации спектра модуляции. Тем самым будет достигнуто уменьшение уровня шума и реверберации. Это свойство модуляционного спектра используется в системе автоматического распознавания речи, позволяя сделать систему более независимой от окружающей акустической обстановки. Для этой цели используются постоянные полосовые *КИХ*-фильтры (1-16 Гц) с различными амплитудно-частотными характеристиками. Также может использоваться модуляционный фильтр, который усиливает модуляционные компоненты в диапазоне 2-8 Гц, для улучшения разборчивости речи или может быть предпринята оценка условий окружающей среды, по результатам которой выбирается или управляется модуляционный фильтр.

Используя свойства речи в модуляционной области, можно выполнить три действия, необходимые для работы детектора в неблагоприятной акустической обстановке: 1) устранить большую часть энергии шума, сконцентрированную в частотном диапазоне 0-1 Гц. Эта операция позволяет сделать метод инвариантным к динамическим фоновым шумам; 2) оценить стандартное отклонение шумовых компонент в модуляционном диапазоне 1-16 Гц. Данная операция позволяет непрерывно отслеживать изменение характеристик окружающего шума более точно, чем стандартными методами; 3) определить весовые коэффициенты, определяющие вероятность наличия в частотных полосах речевых компонент. Данное действие позволяет дополнительно ослабить уровень шума, выделив тем самым речевые компоненты в модуляционном диапазоне 1-16 Гц.

Основное отличие данного метода заключается в том, что параметры окружающей среды оцениваются и при наличии речевых участков. Блок-схема данного метода показана на рис. 1.1.

Речевой сигнал $x(n/f_S)$ разбивается на M равных частотных полос при помощи *ДПФ* модулированного банка полифазных фильтров. Огибающая амплитуды $Y_K(nM/f_S)$ сигнала $X_K(nM/f_S)$ вычисляется в каждой частотной полосе. Затем она поступает на вход блока обработки огибающей амплитуды. На выход данного блока поступают два сигнала: $S_K(nM/f_S)$ – огибающая амплитуды для речевых компонент; $N_K(nM/f_S)$ – огибающая амплитуда для шумовых компонент.

Блок обработки огибающей амплитуды (рис. 1.2) выполняет три основные задачи. Во-первых, удаляет основную часть энергии шумовых компонент, которая сконцентрирована в модуляционном диапазоне 0-1 Гц. Во-вторых, в нем вычисляется огибающая амплитуды шумовых компонент. В-третьих, определяет *SNR* сигнала и вычисляет коэффициент ослабления для каждой частотной полосы.

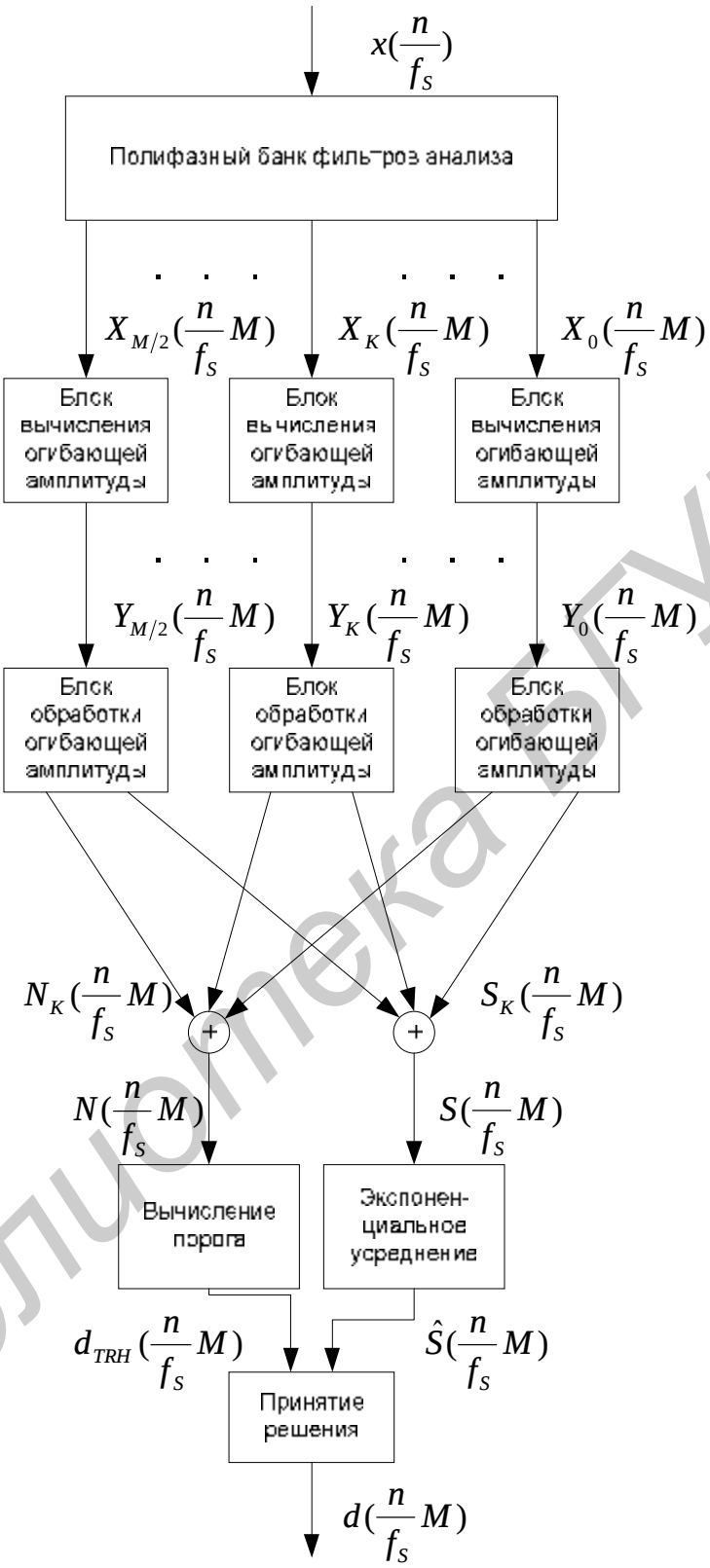

Рис. 1.1. Блок-схема основанного на обработке модуляционного спектра детектора речи

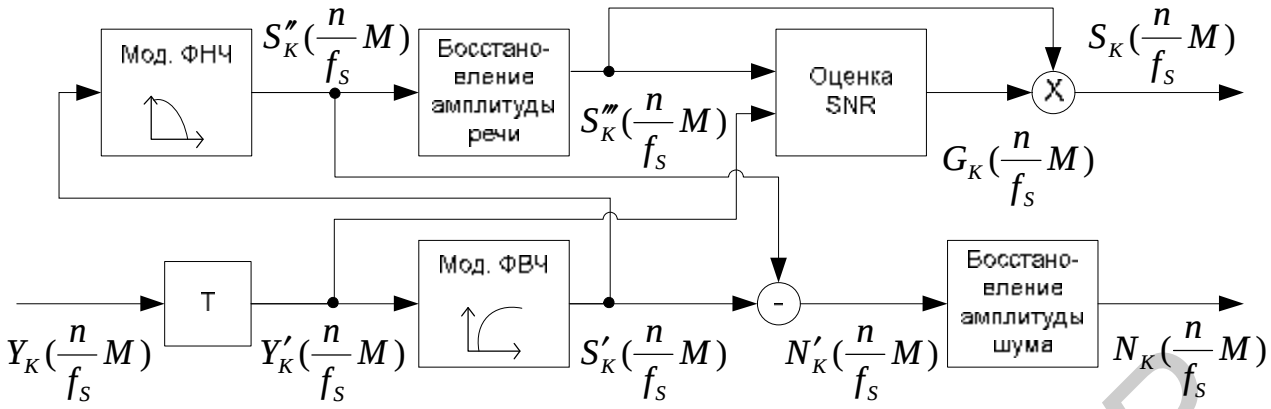

Рис. 1.2. Схема блока обработки огибающей амплитуды

Огибающая амплитуды $Y_K(nM/f_S)$ в каждой частотной полосе преобразуется в нелинейную шкалу согласно

$$Y'_K(nM/f_S) = Y_K(nM/f_S)^{1/g}, \tag{1.3}$$

где g – определяет степень сжатия. Далее, огибающая амплитуды $Y'_K(nM/f_S)$ фильтруется фильтром верхних частот с частотой среза 1 Гц.

Результатом данной операции является устранение основной части энергии шума. Затем сигнал $Y''_K(nM/f_S)$ фильтруется фильтром нижних частот с частотой среза 16 Гц. Данная операция удаляет шумовые компоненты, которые сконцентрированы в модуляционном диапазоне свыше 16 Гц и позволяет оценить огибающую их амплитуды по формуле

$$N'_K(nM/f_S) = S'_K(nM/f_S) - S''_K(nM/f_S). \tag{1.4}$$

Сигнал $N'_K(nM/f_S)$ является приблизительной оценкой огибающей амплитуды шумовых компонент в модуляционном диапазоне 1-16 Гц. Обработка огибающей фильтром верхних частот удаляет речевые компоненты, находящиеся ниже 1 Гц. Это приводит к появлению негативных значений в отфильтрованных огибающих $S''_K(nM/f_S)$ и $N'_K(nM/f_S)$.

Поэтому производится восстановление огибающих речевых и шумовых компонент согласно формулам:

$$S'''_K(\frac{n}{f_S}M) = S''_K(\frac{n}{f_S}M) + std(S'''_K(\frac{n}{f_S}M)), \tag{1.5}$$

$$N(\frac{n}{f_S}M) = N'_K(\frac{n}{f_S}M) + std(N'_K(\frac{n}{f_S}M)), \tag{1.6}$$

где std – стандартное отклонение сигнала, вычисляемое на фрейме длиной $L = f_S/4M$. Длина фрейма L соответствует модуляционной частоте 4 Гц, около которой сконцентрирована основная часть энергии речевого сигнала.

Затем для каждой частотной полосы вычисляется коэффициент усиления по следующей формуле:

$$G(\frac{n}{f_S}M) = (\hat{Y}'_K(\frac{n}{f_S}M) - \hat{S}'''_K(\frac{n}{f_S}M))/\hat{Y}'_K(\frac{n}{f_S}M), \quad (1.7)$$

где $\hat{S}'''_K(nM/f_S)$ и $\hat{Y}'_K(nM/f_S)$ сигналы $S'''_K(nM/f_S)$ и $Y'_K(nM/f_S)$ после экспоненциального усреднения соответственно. Перемножение сигнала $S'''_K(nM/f_S)$ на весовую функцию $G_K(nM/f_S)$ обеспечивает дальнейшее ослабление энергии шума в огибающих речевых компонент (1-16 Гц).

Огибающие амплитуды $S_K(nM/f_S)$, $N_K(nM/f_S)$ для речевых и шумовых компонент (рис. 1.2) суммируются для всех частотных полос (рис. 1.3, 1.4):

$$S(nM/f_S) = \sum_{K=1}^{M/2} S_K(nM/f_S), \; N(nM/f_S) = \sum_{K=1}^{M/2} N_K(nM/f_S). \quad (1.8)$$

Суммированный сигнал $S(nM/f_S)$ сглаживается с помощью экспоненциального усреднения для уменьшения вероятности возникновения изолированных ошибок. Суммированный сигнал $N_K(nM/f_S)$ используется для вычисления порога. Порог d_{thr} вычисляется в течение всего времени работы алгоритма.

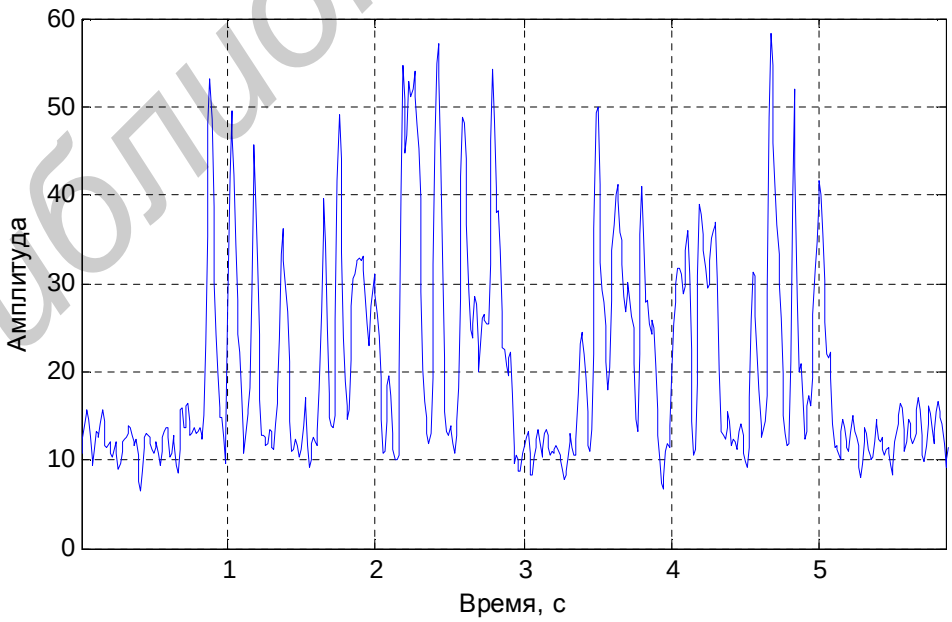

Рис. 1.3. Суммарная огибающая речевых компонент

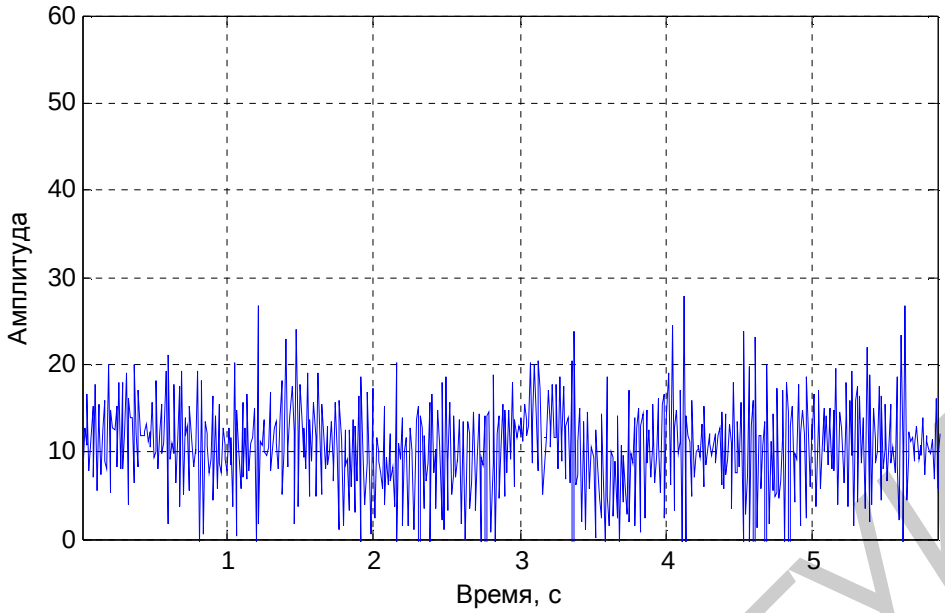

Рис. 1.4. Суммарная огибающая шумовых компонент

Как отмечалось ранее, его оценка вычисляется на основании оценки шумовых компонент в модуляционном диапазоне выше 16 Гц $N(nM / f_S)$ по следующей формуле:

$$d_{thr}(\frac{n}{f_S}M) = mean(N(\frac{n}{f_S}M)) + l \cdot std(N(\frac{n}{f_S}M)), \qquad (1.9)$$

где $std(N(nM / f_S))$ – стандартное отклонение сигнала $N(nM / f_S)$, посчитанное на фрейме длиной $L = f_S / 4M$; $mean(N(nM / f_S))$ – среднее значение сигнала $N_K(nM / f_S)$; l – определяет доверительный предел.

Принятие решения речь/пауза d основано на сравнении сглаженного сигнала суммированных амплитуд $S(nM / f_S)$ с порогом d_{thr} (рис. 1.5).

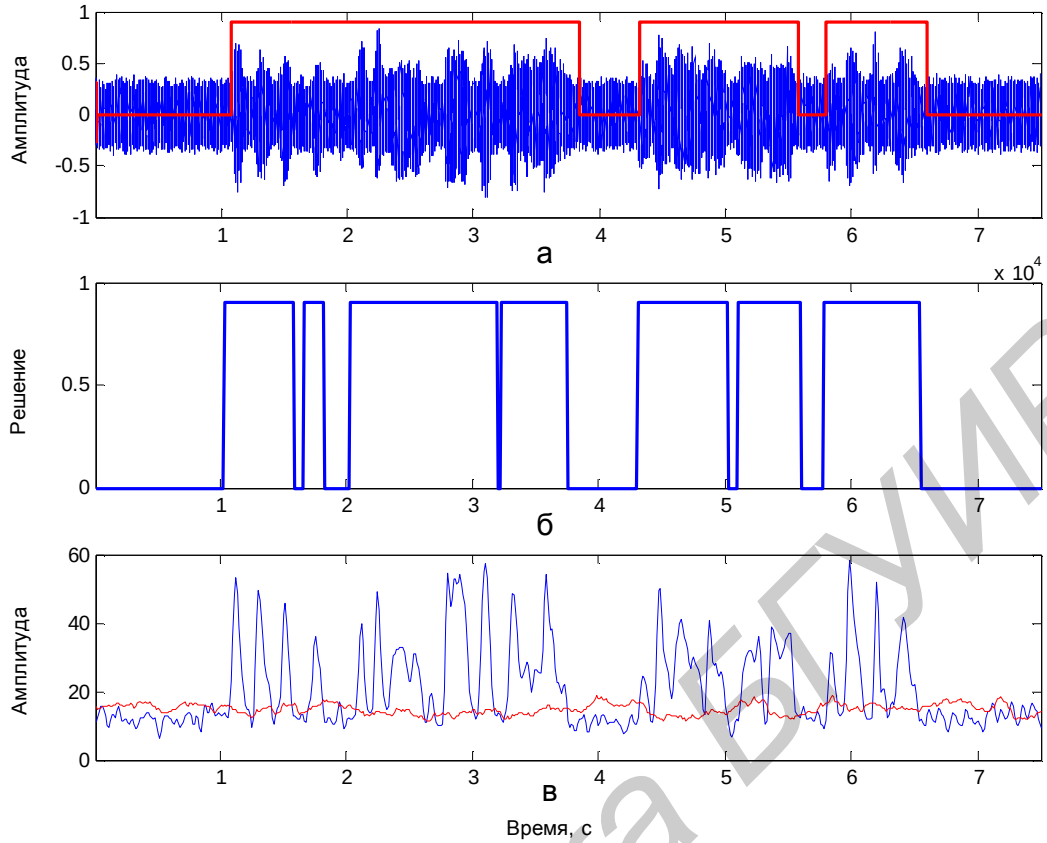

Рис. 1.5. Зашумленный речевой сигнал:
а – эталонное решение речь/пауза; б – решение речь/пауза полученное детектором; в – классификационный параметр и порог

1.4. Критерии оценки детектора речи

Характеристика работы детектора речи может быть оценена следующими объективными параметрами:

$P(S) = N_S /(N_S + N_N)$ – вероятность появления речи; $P(N) = N_N /(N_S + N_N)$ – вероятность появления пауз; $P(A)$ – вероятность правильного определения; $P(B)$ – коэффициент корректности решения речь/пауза, определяемые как

$$P(A) = P(A/S)P(S) + P(A/N)P(N), \qquad (1.10)$$

$$P(B) = P(A/S)P(A/N), \qquad (1.11)$$

где $P(A/S) = N'_S / N_S$ – вероятность правильного определения речи, $P(A/N) = N'_N / N_N$ – вероятность правильного определения пауз.

2. ВЕКТОРНОЕ КВАНТОВАНИЕ

2.1. Определение

Квантование – процесс аппроксимации непрерывных сигналов дискретными значениями, играющий важную роль при сжатии данных или кодировании с целью сокращения числа бит, необходимого для передачи или хранения сигналов с требуемой точностью. Независимое квантование каждого значения сигнала или параметра называется **скалярным квантованием** (СК). Операция квантования здесь сводится к тому, что всем величинам квантуемого параметра, попавшим в некоторый интервал, приписывается одно и то же заданное значение. Скалярное квантование бывает **равномерным** и **неравномерным**. В последнем случае меньшие интервалы используются в областях с большей вероятностью появления квантуемых величин, а большие интервалы – в областях с меньшей вероятностью появления.

Совместное квантование блока параметров называется блочным или **векторным квантованием** (ВК). Векторное квантование представляется как процесс исключения избыточности, в котором эффективно используются четыре взаимосвязанных свойства векторных параметров: линейная зависимость (корреляция), нелинейная зависимость, форма функции плотности вероятности (ФПВ) и многомерность вектора, тогда как при скалярном квантовании можно эффективно использовать только линейную зависимость и форму ФПВ. Нелинейная зависимость играет существенную роль при квантовании спектральных параметров речи, тогда как многомерность важна при квантовании формы сигнала. Вследствие относительно высокой стоимости векторного квантования (зависимость стоимости от размерности вектора и количества бит на координату обычно имеет экспоненциальный характер) на современном уровне развития вычислительной техники преимущества векторного квантования проявляются в основном при скоростях передачи до 1 бита на параметр. Как раз в этой области рабочие характеристики скалярных квантователей резко ухудшаются. Таким образом, к числу *основных преимуществ* данного метода кодирования следует отнести следующие характеристики: а) возможность устранить корреляционные связи между координатами вектора признаков; б) возможность учесть нелинейные зависимости между координатами вектора признаков. Все это позволяет существенно снизить избыточность цифрового представления сигнала, а следовательно, сократить алфавит возможных состояний акустической модели.

2.2. Постановка задачи

Предполагается, что $x = [x_1, x_2, ..., x_N]^T$ представляет собой N-мерный вектор, компоненты которого $\{x_k, 1 \le k \le N\}$ - действительные случайные величины с непрерывным распределением амплитудных значений (T - символ транспонирования). При *ВК* вектор x отображается в N-мерный действительный вектор y с дискретными значениями амплитуд. Этот процесс называется квантованием x в y, а y представляет собой квантованное значение x. Про-

цесс квантования можно записать в виде следующего выражения:

$$y = q(x), \tag{2.1}$$

где $q(\cdot)$ – оператор квантования, y называется также преобразованным вектором или выходным вектором, соответствующим x. Обычно y принимает одно значение из ограниченного множества $Y = \{y_i, 1 \le i \le L\}$, где $y_i = [y_{i1}, y_{i2}, \mathbf{K}, y_{iN}]^T$. Множество Y называется кодовой книгой преобразования или просто кодовой книгой. L – размер кодовой книги, а $\{y_i\}$ – множество кодовых векторов. Размер кодовой книги называют также числом уровней (этот термин пришел из терминологии скалярного квантования). Таким образом, можно говорить о кодовой книге с L уровнями или об L-уровневом квантователе. Для построения такой кодовой книги N-мерное пространство случайного вектора x разделяется на L областей или ячеек $\{C_i, 1 \le i \le L\}$ и с каждой ячейкой C_i связывается вектор y_i. Квантователь назначает кодовый вектор y_i, если x лежит в C_i:

$$q(x) = y_i, \text{ если } x \in C_i. \tag{2.2}$$

Процесс построения кодовой книги известен также как процесс обучения или заполнения кодовой книги. На рис. 2.1 приведен пример разделения двухмерного пространства (N=2) при векторном квантовании. Область, выделенная утолщенными линиями, представляет собой ячейку C_i. Любой входной вектор x, лежащий в ячейке C_i, квантуется как y_i. Положения кодовых векторов других ячеек обозначены точками. Общее число кодовых векторов в приведенном примере составляет L=24. При N=1 векторное квантование вырождается в скалярное квантование. На рис. 2.2 показан пример разделения действительной оси при скалярном квантовании. Значения кода (выходные или преобразованные уровни) отмечены точками. Здесь также любое значение входной величины x, которое лежит в интервале C_i, квантуется как y_i (число уровней квантования L=10). Скалярное квантование имеет особенность, состоящую в том, что хотя размеры ячеек могут быть различными, все они имеют одинаковую форму, а именно, все они представляются интервалами на действительной оси. Для сравнения на рис. 2.1 были показаны двухмерные ячейки, имеющие различную форму. Благодаря возможности построения в многомерном пространстве ячеек различной формы векторное квантование имеет преимущества перед скалярным квантованием.

При квантовании x в y возникает ошибка квантования. Отклонение x от y может быть определено мерой искажения $d(x, y)$, называемой также мерой расхождения или мерой расстояния.

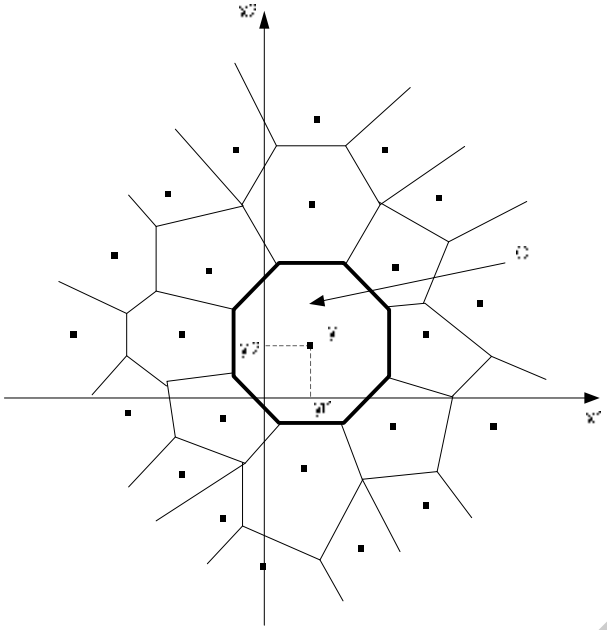

Рис. 2.1. Разделение двухмерного пространства (N=2) на L=24 ячейки (формы разных ячеек могут отличаться весьма значительно)

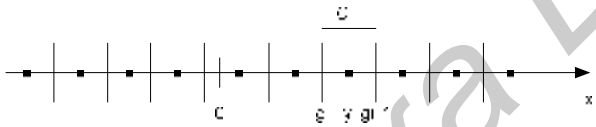

Рис. 2.2. Разбиение действительной прямой на L=10 ячеек (интервалов) при скалярном квантовании

В случае передачи векторов $y(n)$ в различные моменты времени n можно определить результирующее среднее искажение по следующему выражению:

$$D = \lim_{M\to\infty} \frac{1}{M} \sum_{n=1}^{M} d\left[x(n), y(n)\right]. \tag{2.3}$$

Если векторный процесс $x(n)$ обладает свойствами стационарности и эргодичности (эргодичность позволяет заменить среднее по ансамблю на среднее по отсчетам (или по времени)), то выборочное среднее в пределе стремится к математическому ожиданию:

$$\begin{aligned} D = E\left[d(x,y)\right] &= \sum_{i=1}^{L} P(x\in C_i) E\left[d(x,y_i)\middle| x\in C_i\right] = \\ &= \sum_{i=1}^{L} P(x\in C_i) \int_{x\in C_i} d(x,y_i) p(x) dx, \end{aligned} \tag{2.4}$$

где $E[\cdot]$ – математическое ожидание, $P(x\in C_i)$ – вероятность того, что x на-

ходится в C_i, $p(x)$ – многомерная функция плотности вероятности (ФПВ) величины x.

При передаче каждый вектор y_i кодируется двоичным кодовым словом c_i, содержащим B_i бит. В общем случае различные кодовые слова имеют различную длину. При этом скорость передачи T определяется выражениями:

$$T = B \cdot F_C \text{ бит/с}, \quad B = \lim_{M \to \infty} \frac{1}{M} \sum_{n=1}^{M} B(n) \text{ бит/вектор}, \tag{2.5}$$

где $B(n)$ – число бит, использованных для кодирования вектора $x(n)$ в момент времени n, F_C – число кодовых слов, передаваемых за одну секунду.

Полезно также определить среднее число бит на параметр или координату вектора как отношение числа бит, использованных для кодирования вектора, к количеству координат в векторе – N:

$$R = \frac{B}{N} \text{ бит/координата}. \tag{2.6}$$

При использовании кодовой книги размером L максимальное число бит, необходимое для кодирования каждого вектора, можно определить согласно следующему выражению:

$$B_{\max} = \log_2 L. \tag{2.7}$$

Процесс векторного квантования осуществляется следующим образом. Кодер и декодер содержат кодовую книгу, состоящую из векторов (блоков) параметров C_i, представляющих входные векторы параметров x. Для каждого блока входных параметров в кодере реализуется поиск записи C_i, обеспечивающей наилучшее соответствие входному вектору x. Индекс кодового вектора, наилучшим образом представляющего блок входных параметров, передается в декодер, который использует кодовую книгу как таблицу для реконструкции входного сигнала. На рис. 2.3 представлена наиболее общая схема векторного квантования. При построении системы сжатия данных квантователь должен проектироваться таким образом, чтобы для заданной скорости передачи искажения выходного сигнала были минимальны. При проектировании квантователя один из главных вопросов заключается в выборе меры искажений.

2.3. Мера искажений

Мера искажений найдет применение лишь в том случае, если она выражается в удобной математической форме, пригодной для анализа и расчетов, и если она отражает свойства субъективного восприятия, так что различия в величинах искажений характеризуют соответствующие различия качества речи.

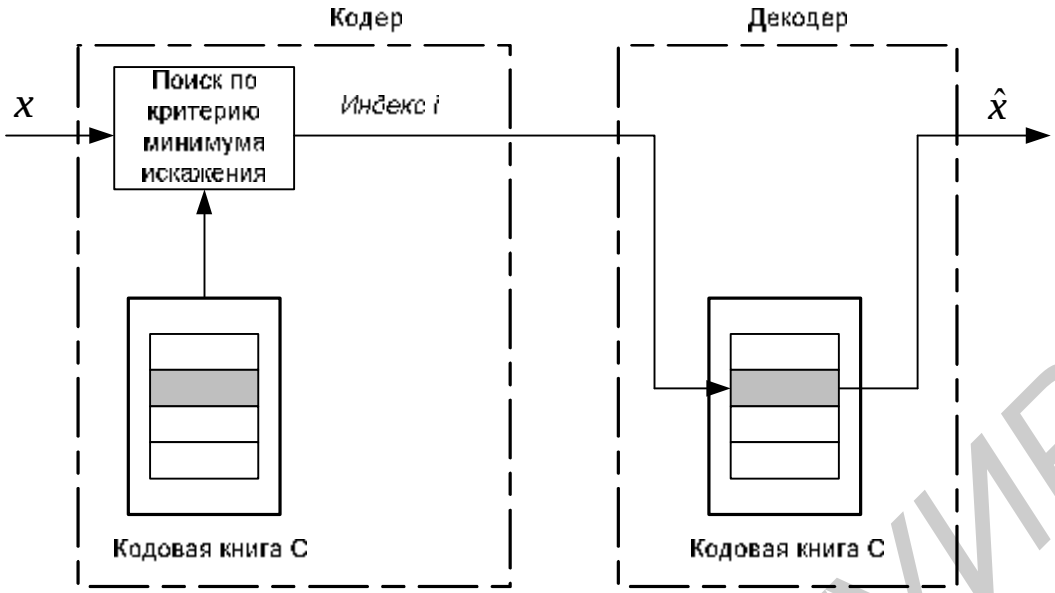

Рис. 2.3. Упрощенная схема процесса векторного квантования

Большинство из используемых в настоящее время мер искажений поддается математическому анализу и отражает некоторые особенности восприятия. Тем не менее в одних случаях уменьшение искажения на несколько децибел весьма заметно на слух, а в других случаях может быть и вовсе незаметно. Несмотря на то, что объективные меры искажений могут служить необходимыми и полезными инструментами при проектировании систем кодирования речи, для принятия правильных решений о выборе путей улучшения характеристик системы необходимы периодические субъективные испытания качества. Рассмотрим наиболее распространенные меры искажений.

1. Среднеквадратичная ошибка или отклонение (СКО) определяется по формуле

$$d_2(x,y)=\frac{1}{N}(x-y)^T(x-y)=\frac{1}{N}\sum_{k=1}^{N}(x_k-y_k)^2, \qquad (2.8)$$

где x, y – анализируемые векторы, N – размерность векторов, x_k, y_k – k-е координаты векторов x, y соответственно.

2. Общая мера искажений основана на критерии L_r, определяемом в виде выражения

$$d_r(x,y)=\frac{1}{N}\sum_{k=1}^{N}\left|x_k-y_k\right|^r. \qquad (2.9)$$

Заметим, что при r=2 общая мера искажения преобразуется в СКО. Часто используются два других значения r=1 и $r=\infty$. Соответственно d_1 представляет собой среднее значение абсолютной ошибки, а d_∞ характеризует максимальную ошибку. При кодировании речи наибольшее распространение получила мера d_2, хотя иногда используются меры d_1 и d_∞.

3. Взвешенная среднеквадратичная ошибка. При использовании меры d_2 предполагается, что искажения, обусловленные квантованием различных параметров $\{x_k\}$, учитываются с одинаковым весом. В общем случае можно ввести неравные веса и тем самым увеличить вклад искажений некоторых параметров в общую меру искажений. Взвешенную СКО можно определить как

$$d_W(x,y)=(x-y)^T W(x-y), \tag{2.10}$$

где W – положительно определенная взвешивающая матрица.

Часто в задачах распознавания образов принимают $W=\Gamma^{-1}$, где Γ – ковариационная матрица случайного вектора x:

$$\Gamma = E\left[(x-\bar{x})(x-\bar{x})^T\right], \tag{2.11}$$

где $\bar{x}$ – среднее по ансамблю реализаций (в нашем случае математическое ожидание). В этом случае d_W можно представить в виде следующего выражения, которое в данном случае известно под названием расстояние Махаланобиса:

$$d_W(x,y)=(x-y)^T\Gamma^{-1}(x-y). \tag{2.12}$$

Если взвешивающая матрица W не только положительно определена, но и симметрична (как в случае расстояния Махаланобиса), ее можно представить в виде произведения

$$W=P^T P, \tag{2.13}$$

где P – нижняя треугольная матрица с единичной главной диагональю. Векторы x и y можно преобразовать в новый набор векторов $\tilde{x}=Px$ и $\tilde{y}=Py$, тогда

$$d_W(x,y)=(Px-Py)^T(Px-Py)=\left(\tilde{x}-\tilde{y}\right)^T\left(\tilde{x}-\tilde{y}\right)=d_2\left(\tilde{x},\tilde{y}\right). \tag{2.14}$$

Таким образом, взвешенная СКО исходных векторов равна СКО преобразованных векторов. Поэтому с целью упрощения вычислений перед векторным квантованием целесообразно произвести преобразование всех данных согласно выражению, представленному выше.

2.4. Формирование кодовой книги

Как упоминалось выше, для построения кодовой книги с L уровнями N-мерное пространство разделяется на L ячеек $\{C_i, 1\le i\le L\}$ и каждой ячейке C_i приписывается вектор y. В квантователе назначается кодовый вектор y_i, если x лежит в ячейке C_i. Квантователь называется оптимальным (обеспечивающим минимальные искажения), если принятая мера искажения минимизи-

рована по всем L-уровневым квантователям. Существуют два необходимых условия оптимальности.

Первое условие заключается в том, что в оптимальном квантователе используется правило выбора по минимуму искажений, т.е. производится выбор ближайшей ячейки $q(x) = y_i$ тогда и только тогда, когда выполняется условие

$$d(x, y_i) \le d(x, y_j), \quad j \ne i, \quad 1 \le j \le L. \tag{2.15}$$

Другими словами, в квантователе осуществляется выбор кодового вектора, обеспечивающего минимальные искажения x. (При равенстве искажений, обеспечиваемых разными кодовыми векторами, назначается некоторое правило, позволяющее разрешить эту неопределенность и осуществить выбор.)

Второе необходимое условие оптимальности состоит в том, что каждый кодовый вектор y_i выбирается из условия минимизации среднего искажения в ячейке C_i. Таким образом, y_i есть вектор y, минимизирующий выражение

$$D_i = E\left[d(x, y) \middle| x \in C_i\right] = \int_{x \in C_i} d(x, y) p(x) dx. \tag{2.16}$$

Такой вектор называется ***центроидом*** ячейки C_i и записывается в виде

$$y_i = cent(C_i). \tag{2.17}$$

Для расчета ***центроида*** необходимо задать меру искажения. (Ячейки, определяемые таким образом, известны как ячейки ближайших соседей, ячейки Вороного или области Дирихле).

На практике задается набор обучающих векторов $\{x(n), 1 \le n \le M\}$. Некоторое подмножество M_i этих векторов попадает в ячейку C_i. Среднее искажение D_i определяется по следующей формуле:

$$D_i = \frac{1}{M_i} \sum_{x \in C_i} d(x, y_i). \tag{2.18}$$

Для критерия СКО или взвешенной СКО можно показать, что D_i минимизируется при определении ***центроида***:

$$y_i = \frac{1}{M_i} \sum_{x \in C_i} x(n), \tag{2.19}$$

т.е. y_i представляет собой выборочное среднее обучающих векторов, содержащихся в ячейке C_i.

Поскольку решение задачи оптимального квантования в виде конечного выражения неизвестно, данная проблема решается путем итеративного улучшения заданного векторного квантователя. Однако, прежде чем улучшать кван-

тователь, следует его инициализировать. Существует несколько примерно равноценных методов инициализации, рассмотрим один из них.

Шаг 1. Рассчитать центроид всего тренировочного множества.

Шаг 2. Рассчитать невзвешенное евклидово расстояние между центроидом и каждым вектором тренировочного множества.

Шаг 3. Выбрать вектор с максимальным расстоянием в качестве эталона.

Шаг 4. Рассчитать невзвешенное евклидово расстояние между эталонным вектором и каждым вектором тренировочного множества.

Шаг 5. Пусть *M/L* – отношение размера тренировочного множества к размерности кодовой книги. Найти *M/L* векторов, ближайших к эталонному вектору, и рассчитать центроид (вектор начальной кодовой книги) для этой группы векторов.

Шаг 6. Уменьшить тренировочное множество путем удаления из него группы векторов, найденных на предыдущем шаге.

Шаг 7. Повторять шаги со 2-го по 6-й до тех пор, пока в тренировочном множестве не останется векторов.

Рассмотрим один из наиболее распространенных методов построения (тренировки) кодовой книги – **алгоритм К-средних** (его разновидностью является обобщенный алгоритм Ллойда (GLA) или алгоритм Линде-Бьюзо-Грэя (LBG)). В нашем случае $K = L$. Алгоритм разделяет набор обучающих векторов $\{x(n), 1 \le n \le M\}$ на L кластеров (ячеек) C_i таким образом, что удовлетворяются два необходимых условия оптимальности (2.15) и (2.16). Если обозначить через m номер итерации, а через $C_i(m)$ – i-й кластер на m-й итерации с центроидом $c_i(m)$, то алгоритм К-средних можно записать в следующем виде:

Шаг 1. Задание начальных значений. Положим m=0. Выберем тем или иным подходящим методом набор начальных кодовых векторов $c_i(0)$, $1 \le i \le L$.

Шаг 2. Классификация. Классифицируем набор обучающих векторов $\{x(n), 1 \le n \le M\}$ по кластерам C_i с помощью правила “ближайшего соседа”: $x \in C_i(m)$ тогда и только тогда, когда $d[x, c_i(m)] \le d[x, c_j(m)]$ для всех $j \ne i$.

Шаг 3. Коррекция кодового вектора. $m \leftarrow m+1$. Произвести коррекцию кодовых векторов всех кластеров путем вычисления центроидов обучающих векторов каждого кластера $c_i(m) = cent(C_i(m))$, $1 \le i \le L$.

Шаг 4. Проверка на окончание процедуры. Если уменьшение величины общего искажения $D(m)$ на итерации m относительно $D(m-1)$ меньше некоторого порога, процедура заканчивается. В противном случае переход на шаг 2.

Введем понятие **тренировочного отношения**, которое определяется как отношение размера тренировочной базы данных (обучающего множества векторов) к размеру кодовой книги. Для построения качественной кодовой книги тренировочное отношение должно быть в пределах от 10 до 50. До сих пор мы рассматривали построение кодовых книг без внутренней структуры. Использо-

вание их в настоящее время невыгодно, т.к. ведет к длительным процессам тренировки и поиска в них либо к дополнительным затратам памяти.

Структурированные кодовые книги могут использовать быстрые методы поиска и занимают значительно меньший объем. В настоящее время разработано большое количество вариантов структурирования кодовых книг: квантование с расщеплением вектора (Split VQ – Split Vector Quantization), многоуровневое векторное квантование (MSVQ – Multistage Vector Quantization), векторное квантование с древовидной структурой (TSVQ – Tree-Structured Vector Quantization), решетчатое квантование (lattice quantization) и т.д. Рассмотрим подробнее структурированные кодовые книги на примере подходов SVQ и MSVQ.

2.5. Векторное квантование с расщеплением

В ВК с расщеплением входной вектор $x = \left[x_1, x_2, \mathbf{K}, x_p\right]^T \in R^p$ расщепляется или делится на R субвекторов меньшей размерности $x = \left[x^{(1)} x^{(2)} x^{(3)}\ \mathbf{K}\ x^{(R)}\right]^T$. Таким образом, i-й субвектор $x^{(i)}$ имеет размерность d_i, при этом $p = d_1 + d_2 + \mathbf{L} + d_R$.

$$x^{(1)} = \left[x_1 x_2\ \mathbf{L}\ x_{d_1}\right]^T;\ \ x^{(2)} = \left[x_{d_1+1} x_{d_1+2}\ \mathbf{L}\ x_{d_1+d_2}\right]^T;$$
$$x^{(3)} = \left[x_{d_1+d_2+1} x_{d_1+d_2+2}\ \mathbf{L}\ x_{d_1+d_2+d_3}\right]^T. \qquad (2.20)$$

То есть имеем R квантователей, по одному на каждый субвектор. Субвекторы $x^{(i)}$ индивидуально квантуются в $y_k^{(i)}$, при этом входной вектор x квантуется в $y = \left[y_k^{(1)} y_k^{(2)} y_k^{(3)}\ \mathbf{K}\ y_k^{(R)}\right]^T \in R^p$. Квантователи проектируются (обучаются) путем использования соответствующих субвекторов в обучающем множестве. Частный случай ВК с расщеплением: $R = p$, $d_1 = d_2 = \mathbf{L} = d_p = 1$ – получается скалярный квантователь. Предположим, размерность $p = 10$. Полный 30-битный векторный квантователь будет иметь кодовую книгу из 2^{30} кодовых слов. В эквивалентном векторном квантовании с расщеплением из $R = 3$ расщеплений (частей) используются субвекторы с размерностями $d_1 = 3, d_2 = 3, d_3 = 4$, с каждым субвектором ассоциируется 10-битная кодовая книга, имеющая 2^{10} кодовых слов. Расщепленное ВК возможно, если

$$d(x, y) = \sum_{i=1}^{R} d\left(x^{(i)}, y^{(i)}\right). \qquad (2.21)$$

Данное свойство истинно для L_r дистанции и для взвешенной L_2 дистанции, если матрица весов W является диагональной.

2.6. Многоуровневое векторное квантование

В многоуровневом ВК, включающем R уровней, используются R квантователей $Q_1, Q_2, \mathbf{K}, Q_R$. Соответствующие кодовые книги обозначаются как $C_1, C_2, \mathbf{K}, C_R$. Размеры этих кодовых книг - $N_1, N_2, \mathbf{K}, N_R$. Суммарный размер кодовых книг определяется как $N = N_1 + N_2 + \mathbf{L} + N_R$. Записи i-й кодовой книги C_i – $y_1^{(i)}, y_2^{(i)}, \mathbf{K}, y_N^{(i)}$. На рис. 2.4 показана структура всей системы.

Процедура многоуровневого квантования следующая. Входной вектор x сначала квантуется при помощи квантователя Q_1 в $y_k^{(1)}$. Ошибка квантования $e_1 = x - y_k^{(1)}$ далее квантуется Q_2 в $y_k^{(2)}$. Ошибка второго уровня определяется как $e_2 = e_1 - y_k^{(2)}$ и квантуется на третьем уровне. Процесс повторяется, и на R-м уровне ошибка e_{R-1} квантуется Q_R в $y_k^{(R)}$, в результате на выходе схемы имеем ошибку e_R.

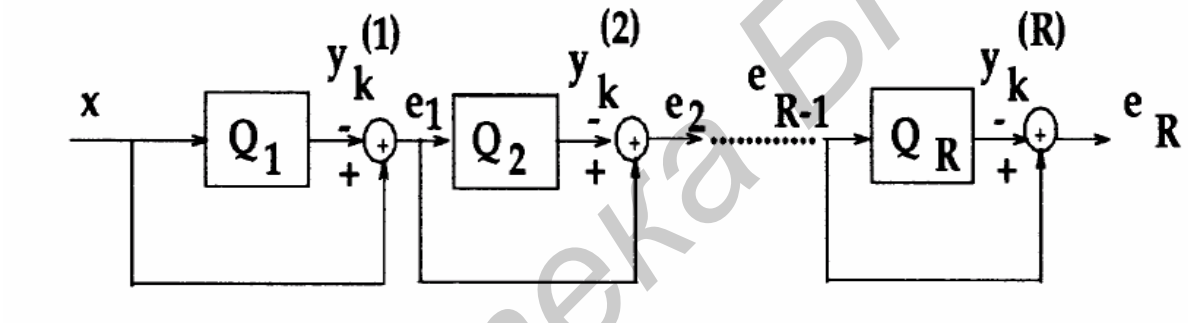

Рис. 2.4. Схема многоуровневого векторного квантования

Таким образом, оригинальный вектор x квантуется в $y = y_k^{(1)} + y_k^{(2)} + \mathbf{L} + y_k^{(R)}$. Общая ошибка квантования определяется как $x - y = e_R$. Уменьшение требуемой памяти и сложности поиска проиллюстрируем на простом примере. Полный (одинарный) 30-битный векторный квантователь будет иметь одну кодовую книгу 2^{30} кодовых векторов (не может использоваться на практике из-за большого объема требуемой памяти). Эквивалентный многоуровневый *ВК* с $R = 3$ уровнями будет иметь три 10-битные кодовые книги C_1, C_2, C_3. Общее количество векторов, хранящихся в памяти, будет составлять $3 \cdot 2^{10} = 3072$, что практически реализуемо. Следовательно, сложность поиска также существенно снизится.

Обучение происходит, как было описано выше, только набор ошибок предыдущего уровня используется в качестве обучающего множества последующего уровня.

3. ВЕКТОРНОЕ КВАНТОВАНИЕ КОЭФФИЦИЕНТОВ ЛИНЕЙНОГО ПРЕДСКАЗАНИЯ

3.1. Линейное предсказание речи

Линейное предсказание (*LPC* – Linear Predictive Coding) является одним из наиболее эффективных методов анализа речевого сигнала и применяется во многих кодерах речи. Этот метод становится доминирующим при сокращённом представлении речи с целью её низкоскоростной передачи и экономного хранения. Важность метода обусловлена высокой точностью получаемых оценок и относительной простотой вычислений.

Основной принцип метода линейного предсказания состоит в том, что текущий отсчёт речевого сигнала можно аппроксимировать линейной комбинацией предшествующих отсчётов. Коэффициенты предсказания при этом определяются однозначно минимизацией среднего квадрата разности между отсчётами речевого сигнала и их предсказанными значениями (на конечном интервале). Коэффициенты предсказания – это весовые коэффициенты, используемые в линейной комбинации.

Основные положения метода линейного предсказания хорошо согласуются с моделью речеобразования, в которой речевой сигнал представлен в виде сигнала на выходе линейной системы с переменными во времени параметрами, возбуждаемой квазипериодическими импульсами (в пределах вокализованного сегмента) или случайным шумом (на невокализованном сегменте). Метод линейного предсказания позволяет точно и надёжно оценить параметры этой линейной системы с переменными коэффициентами.

Общий спектр, обусловленный излучением, речевым трактом и возбуждением, описывается с помощью линейной системы с переменными параметрами и передаточной функцией, выраженной следующей формулой:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}, \qquad (3.1)$$

где $\{a_k\}$ – множество коэффициентов, которые называются *LPC*-параметрами, или коэффициентами предсказателя; p – порядок предсказателя (количество коэффициентов); G – коэффициент усиления.

Основная задача анализа на основе линейного предсказания заключается в непосредственном определении параметров $\{a_k\}$ по речевому сигналу с целью получения хороших оценок его спектральных свойств путём использования уравнения (3.1). Вследствие изменения свойств речевого сигнала во времени коэффициенты линейного предсказания должны оцениваться на коротких сегментах речи. Основным подходом является определение параметров линейного предсказания таким образом, чтобы минимизировать дисперсию погрешности на коротком сегменте сигнала.

Как уже было сказано выше, очередной отсчет речевого сигнала может быть аппроксимирован линейной комбинацией предыдущих отсчетов, т.е.

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k), \tag{3.2}$$

где $\tilde{s}(n)$ – предсказанное значение речевого отсчета, $s(n)$ – оригинальный отсчет речевого сигнала.

Таким образом, ошибку предсказания можно представить как разность между оригинальным значением и предсказанным:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{3.3}$$

Взяв z-преобразование от выражения (3.3), мы получим

$$E(z) = S(z)A(z), \tag{3.4}$$

где $A(z)$ - это инверсия от $H(z)$ в выражении (3.1), т.е. $A(z)$ – *инверсный фильтр*, уравнение которого может быть записано в следующем виде:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}. \tag{3.5}$$

Из-за времязависимой природы речевого сигнала коэффициенты предсказания должны определяться на коротких сегментах речи (10 – 30 мс).

Таким образом, необходимо определить множество коэффициентов предсказателя, которые бы минимизировали ошибку на всем сегменте сигнала. Полученные параметры будут являться параметрами системной функции $H(z)$ модели речеобразования в выражении (3.1).

Средняя ошибка кратковременного предсказания определяется по следующему выражению:

$$E = \sum_{n} e^2(n) = \sum_{n} \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2. \tag{3.6}$$

Для определения значений $\{a_k\}$, минимизирующих ошибку E, необходимо взять частную производную по всем коэффициентам и приравнять ее к нулю:

$$\frac{\partial E}{\partial a_i} = -2 \sum_{n} \left\{ \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right] s(n-i) \right\} = 0, \tag{3.7}$$

что дает следующее уравнение:

$$\sum_n s(n)s(n-i) = \sum_n \sum_{k=1}^{p} a_k s(n-k)s(n-i). \qquad (3.8)$$

Если изменить порядок суммирования в правой части уравнения (3.8), то получим следующее выражение:

$$\sum_n s(n)s(n-i) = \sum_{k=1}^{p} a_k \sum_n s(n-k)s(n-i),\ i = 1,..., p. \qquad (3.9)$$

Если сделать замену:

$$f(i,k) = \sum_n s(n-i)s(n-k), \qquad (3.10)$$

то уравнение (3.9) может быть записано в следующем виде:

$$\sum_{k=1}^{p} a_k f(i,k) = f(i,0),\ i = 1,..., p. \qquad (3.11)$$

Эта система из p уравнений с p неизвестными может быть эффективно решена относительно неизвестных коэффициентов $\{a_k\}$.

В настоящее время применяется множество методов *LP*-анализа, но наибольшее распространение получили методы на основе автокорреляции и автоковариации, при этом первый имеет меньшую вычислительную сложность и всегда обеспечивает синтез стабильного фильтра-предсказателя.

3.2. Векторное квантование LPC-параметров (LSF-коэффициентов)

Для речевых сегментов с высокой частотой основного тона *LP*-анализ получает фильтр-синтезатор с узкими областями спектрального резонанса. Для решения этой проблемы применяется расширение частотной полосы, которое позволяет уменьшить остроту этих пиков, а также решает некоторые проблемы числовой точности, связанные с близостью полюсов к единичной окружности. Количественно величина расширения (в герцах) определяется следующим образом:

$$\Delta B = -\frac{1}{pF_S}\ln(g), \qquad (3.12)$$

где F_s – частота дискретизации (Гц).

Далее проводится модификация *LPC*-коэффициентов:

$$a_k^{'} = a_k g^k,\ \ 1 \le k \le p. \qquad (3.13)$$

Квантование *LPC*-параметров является одним из важнейших аспектов *LP*-анализа, поскольку минимизация ёмкости кодирования является основной целью в приложениях обработки речи. Несмотря на то, что можно и непосред-

ственно квантовать параметры предсказания, такой подход требует высокой точности представления (8-10 бит на параметр). Это связано с тем, что малые изменения параметров предсказания приводят к большим изменениям в расположении полюсов и, следовательно, к возможной неустойчивости синтезирующего *LP*-фильтра. Поэтому непосредственное квантование параметров не нашло применения.

Для квантования обычно используют альтернативное представление параметров линейного предсказания – **линейные спектральные пары (частоты)** (*LSP* – Line Spectral Pairs или *LSF* – Line Spectral Frequencies).

Чтобы получить *LSF*-коэффициенты, p нулей функции $A_p(z)$ отражаются на единичную окружность посредством двух z-преобразований $P(z)$ и $Q(z)$ $(p+1)$-го порядка:

$$P_{p+1}(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}); \tag{3.14}$$

$$Q_{p+1}(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}). \tag{3.15}$$

Из этого следует, что

$$A_p(z) = \frac{1}{2}\left[P_{p+1}(z) + Q_{p+1}(z)\right]. \tag{3.16}$$

LSF-коэффициенты представляют собой угловые позиции корней $P(z)$ и $Q(z)$ на единичной окружности в диапазоне $0 \le w_i \le p$. Они имеют следующие свойства:

- все корни $P(z)$ и $Q(z)$ лежат на единичной окружности;
- корни чередуются на единичной окружности, т.е. выполняется следующее неравенство:

$$0 \le w_{q,0} \le w_{p,0} \le w_{q,1} \le w_{p,1} \mathbf{K}, \le p \ . \tag{3.17}$$

При выполнении данных условий синтезирующий фильтр $H(z)$ является стабильным. *LSF* обладают следующими важными свойствами:

- расстояние между *LSF*-коэффициентами определяет амплитуду спектральной плотности мощности (рис. 3.1);
- блок из двух или трех близко расположенных *LSF* сигнализирует о наличии максимума в спектре мощности (соответствует формантной частоте), в то время как расположенные с большим промежутком *LSF* соответствуют минимуму (см. рис. 3.1);
- в общем случае спектральная чувствительность каждого LSF локализована, т.е. при небольшом изменении одного из LSF спектр изменится только в окрестности этого *LSF*-параметра (рис. 3.2).

Итак, для получения параметров *LSF* следует провести *LP*-анализ над взвешенными сегментами речевого сигнала и далее трансформировать полученные *LPC*-коэффициенты.

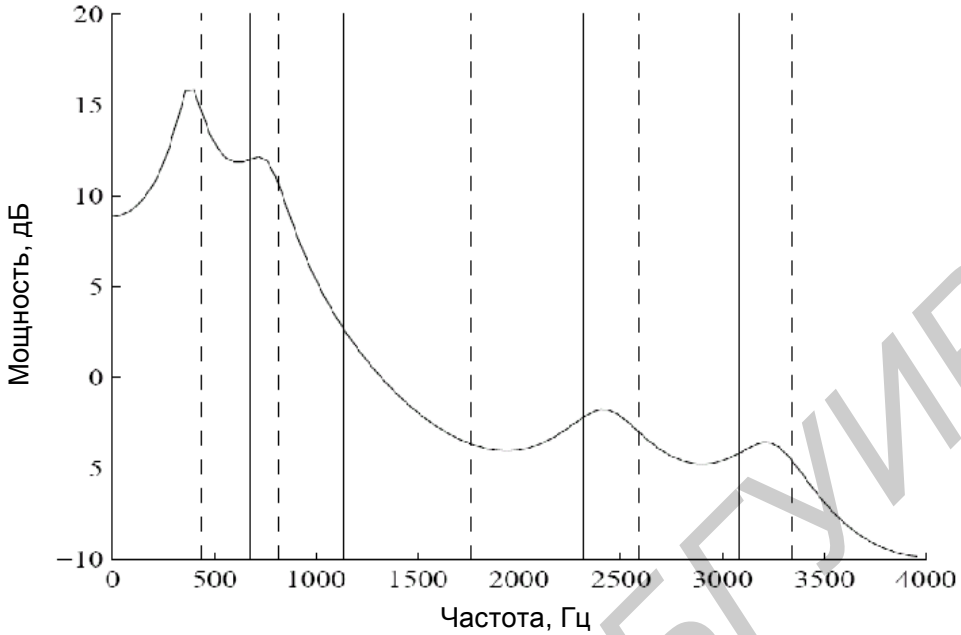

Рис. 3.1. LPC-спектр мощности с соответствующими *LSF* (вертикальные линии)

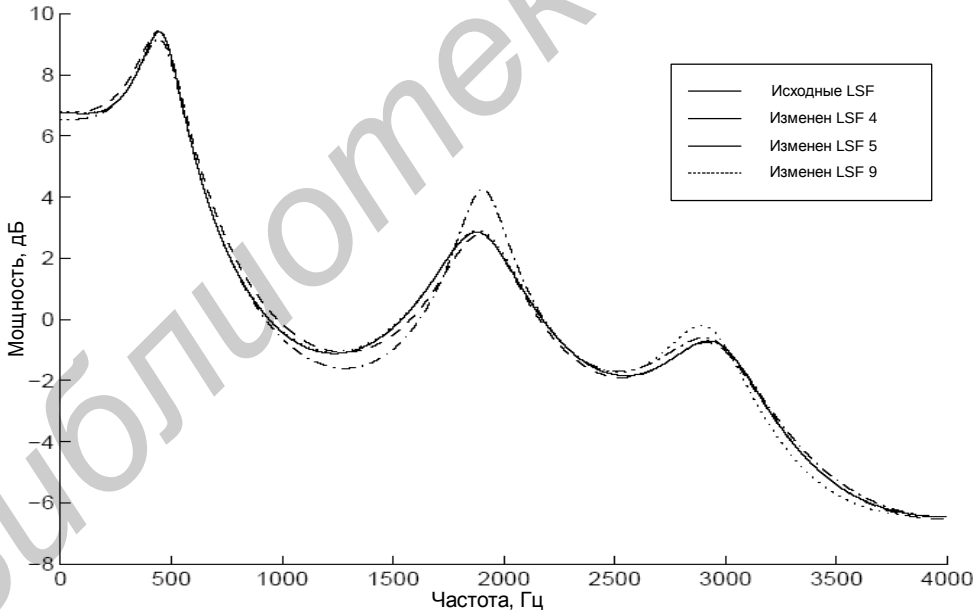

Рис. 3.2. Влияние локального изменения LSF на LPC-спектр мощности. LSF 4 был изменен с 1315 до 1275 Гц, LSF 5 – с 1745 до 1800 Гц, LSF 9 – с 3025 до 2995 Гц

Наиболее распространенные меры искажений, применяемые при построении квантователя в вокодерах, использующих линейное предсказание, представлены ниже.

1. Логарифмическое искажение спектра. Данная мера искажения используется при оценке качества квантователя каких-либо спектральных характеристик. Рассмотрим спектры мощности $S(w)$ и $\hat{S}(w)$, соответствующие оригинальному синтезирующему фильтру и модифицированному синтезирующему фильтру (с квантованными коэффициентами). Искажение спектра на фрейме определяется следующим образом:

$$sd = \sqrt{\frac{1}{p}\int_0^p [10\log_{10} S(w) - 10\log_{10} \hat{S}(w)]^2 dw}\,. \qquad (3.18)$$

В свою очередь, $S(w)$ и $\hat{S}(w)$ могут быть определены как

$$S(\omega) = \frac{1}{\left|A(\omega)\right|^2}, \qquad (3.19)$$

$$\hat{S}(\omega) = \frac{1}{\left|\hat{A}(\omega)\right|^2}, \qquad (3.20)$$

где

$$\left|A(w)\right|^2 = \left|1 + \sum_{k=1}^{p} a_k e^{-jwk}\right|^2 \qquad (3.21)$$

является спектром мощности инверсного фильтра. Из вышесказанного следует, что

$$sd = \sqrt{\frac{1}{p}\int_0^p \left|10\log_{10}\frac{\left|\hat{A}(w)\right|^2}{\left|A(w)\right|^2}\right|^2 dw}\,. \qquad (3.22)$$

2. Взвешенное евклидово расстояние. Данная величина служит математическим критерием для минимизации на стадии квантования, поскольку логарифмическое искажение спектра громоздко для вычисления в реальном времени. Доказано, что взвешенное евклидово расстояние эквивалентно логарифмическому искажению спектра при высокой скорости квантования.

Измерение данной величины можно производить непосредственно в *LSF*-области, поскольку *LSF*-коэффициенты хорошо соответствуют спектральной форме. Для выделения отдельной части спектра *LSF*-коэффициентам, соответствующим этой части, может быть назначен

больший вес. Пусть f и $\hat{f}$ – оригинальный и измененный векторы *LSF* соответственно, тогда их евклидово расстояние $d(f, \hat{f})$ будет определяться следующим образом:

$$d(f, \hat{f}) = \left\| f - \hat{f} \right\|^2 . \tag{3.23}$$

При использовании LP-анализа порядка p получаем следующее:

$$d(f, \hat{f}) = \sum_{i=1}^{p} (f_i - \hat{f}_i)^2 . \tag{3.24}$$

Взвешенное евклидово расстояние отличается использованием весов:

$$d(f, \hat{f}) = \sum_{i=1}^{p} [c_i w_i (f_i - \hat{f}_i)]^2 , \tag{3.25}$$

где c_i и w_i – веса для i-го *LSF*-коэффициента. Для фильтра-предсказателя 10-го порядка фиксированные веса c_i определяются следующим образом:

$$c_i = \begin{cases} 1.0, & 1 \le i \le 8; \\ 0.8, & i = 9; \\ 0.4, & i = 10. \end{cases} \tag{3.26}$$

Человеческое ухо не способно различить разницу на высоких частотах с такой же точностью, как на низких частотах. Таким образом, эти веса используются для того, чтобы усилить значение низких частот. Адаптивные веса w_i используются для выделения областей спектральной огибающей с большей энергией (формант). Веса w_i определяются как

$$w_i = \left[S(e^{jw_i}) \right]^r , \tag{3.27}$$

где r – эмпирическая константа.

Для упрощения можно использовать следующую схему вычисления адаптивных весов:

$$w_i = \frac{1}{f_i - f_{i-1}} + \frac{1}{f_{i+1} - f_i} , \tag{3.28}$$

где f_i – *LSF*-коэффициенты (в радианах), $w_0 = 0$, $w_0 + 1 = \pi$.

4. КВАНТОВАНИЕ ПАРАМЕТРОВ В СИНУСОИДАЛЬНОМ ВОКОДЕРЕ С АНТРОПОМОРФИЧЕСКОЙ ОБРАБОТКОЙ РЕЧЕВОГО СИГНАЛА

4.1. Синусоидальный вокодер с антропоморфической обработкой речевого сигнала

В данной вокодерной системе используется подход, согласно которому речь, как на вокализованных, так и на невокализованных участках, представляется в виде набора синусоидальных компонент.

В процессе анализа в исходном фрагменте речевого сигнала выделяется несколько наиболее “важных” для человеческого слуха синусоидальных компонент. Анализ речевого сигнала в кодере можно представить следующим образом – рис. 4.1.

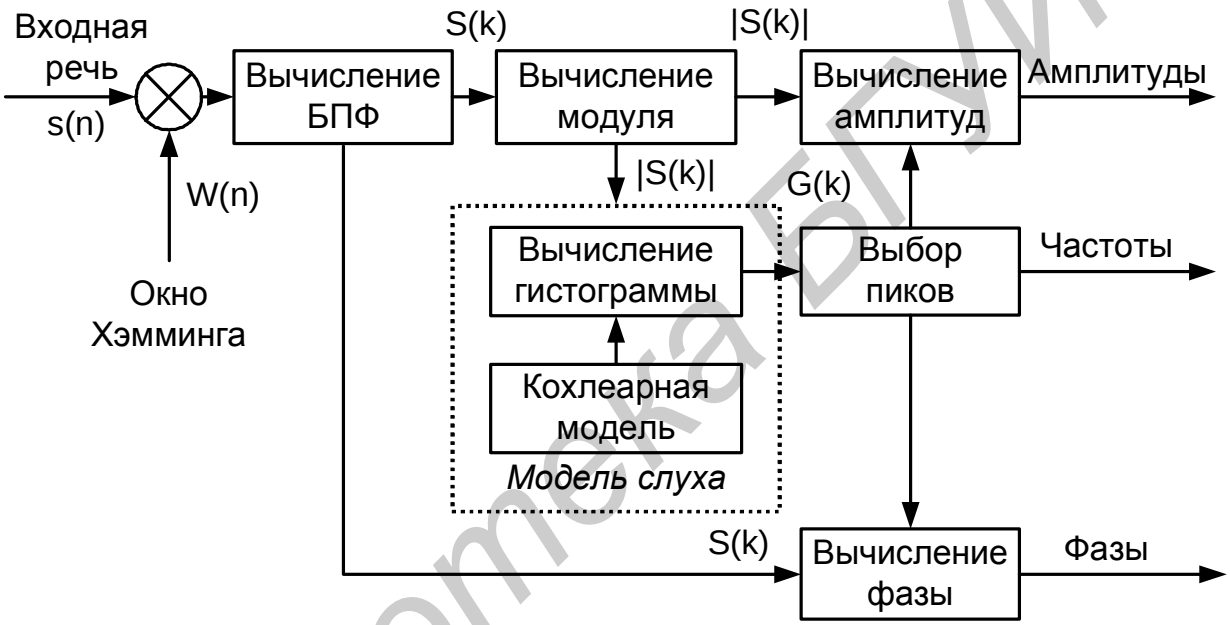

Рис. 4.1. Анализ речевого сигнала в кодере

Входной оцифрованный речевой сигнал $s(n)$ анализируется с помощью спектрального анализа на основе быстрого преобразования Фурье, совмещённого с моделью слуха человека. Выборка очередного фрейма входного оцифрованного с частотой дискретизации F_S речевого сигнала $s(n)$ осуществляется с использованием временного окна Хэмминга $W(n)$. Далее вычисляется дискретное преобразование Фурье $S(k)$ и осуществляется поиск спектральных пиков. В данном случае спектральными пиками считаются точки, значения в которых больше двух ближайших соседей. После этого выполняется процедура отбора спектральных пиков. Отбор производится по спектру сигнала с использованием двух моделей слуха человека. Первая модель представляет собой банк полосовых фильтров, которые имитируют функционирование улитки человеческого уха. Вторая модель представляет систему слуха человека на уровне слухового нерва. Она базируется на вычислении так называемой групповой интервальной гистограммы *EIH*, с её помощью можно дифференцировать присутствующие

частотные составляющие обрабатываемого речевого сигнала по их “важности” и информативности для человеческого восприятия.

Процесс синтеза речи в данном вокодере сводится к суммированию сгенерированных синусоидальных компонентов с найденными в процессе анализа амплитудами, фазами и частотами (рис. 4.2).

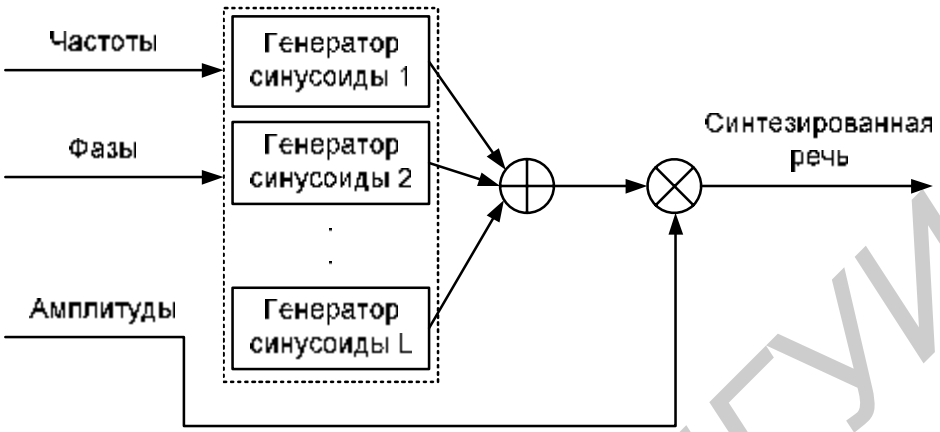

Рис. 4.2. Схема синтеза речевого сигнала

Однако, поскольку параметры речевого сигнала изменяются во времени, а сам сигнал представляется на отдельно взятых фреймах ограниченным (не бесконечным) набором синусоид с определёнными частотами, амплитудами и фазами, то на границах синтезируемых фреймов генерируемый речевой сигнал будет меняться скачкообразно. Это приводит к возникновению хриплости и появлению различных искажений в синтезированной речи. Для устранения этих нежелательных эффектов применяются следующие подходы: 1) *метод наложения со сложением* (overlap-add), при котором каждый генерируемый речевой фрейм определённым образом накладывается на предыдущий фрагмент; 2) метод синтеза *с использованием межфреймовой интерполяции* синусоидальных параметров и определением частотных треков.

Результаты моделирования данной вокодерной системы на основе антропоморфической обработки речи показывают, что можно получить восстановленный речевой сигнал с довольно высокой степенью разборчивости и хорошей узнаваемостью диктора даже при ограниченном числе синусоидальных компонент, т.е. количество отобранных наиболее “критичных” для слуха человека частотных составляющих обычно не превышает 8-10.

4.2. Квантование параметров синусоидальной модели

Для передачи по линии связи найденные в процессе анализа параметры должны быть соответствующим образом заквантованы и закодированы. Ставится задача – оптимальным образом заквантовать параметры отобранных синусоидальных составляющих, т.е. данные параметры должны быть представлены минимальным количеством бит, необходимым для сохранения хорошего качества синтезированного сигнала. При этом должны учитываться следующие

особенности речи:

- амплитудный диапазон речевого сигнала составляет около 60 дБ, т.е. кодирование амплитуды наиболее целесообразно проводить в логарифмическом масштабе;
- учитывая свойства речевого сигнала, его частотный диапазон можно ограничить от 20-40 до 3800-3900 Гц (так называемая узкополосная речь телефонного качества);
- в процессе анализа речевой информации в кодере амплитуды и частоты определяются на основе одной и той же характеристики (спектр речевого сигнала);
- общее количество речевых параметров, которые необходимо заквантовать, довольно велико и при числе синусоидальных компонент 5-10 составляет от 15 до 30.

Экспериментально было доказано, что фазы имеют равномерное распределение и достаточно хорошо квантуются, используя скалярное квантование. Поскольку частоты и амплитуды определяются по спектральной характеристике речевого сигнала, для их кодирования используется векторное квантование в пространстве амплитуда–частота (рис. 4.3). При таком подходе в декодер в качестве параметра вместо непосредственных значений амплитуды и частоты передаётся только индекс найденного элемента в кодовой книге.

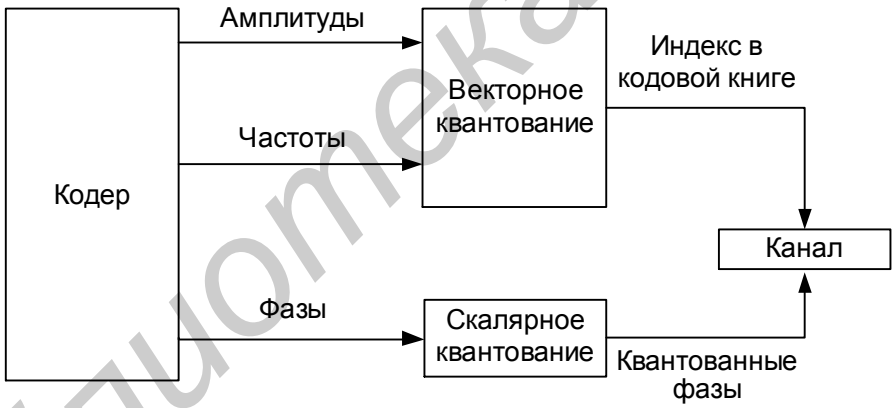

Рис. 4.3. Схема квантования синусоидальных параметров

В рассматриваемой вокодерной системе амплитуды являются целыми числами и лежат в диапазоне от -32768 до 32768 (16 бит). В этом случае для кодирования амплитуды требуется 15 бит плюс один бит для знака. Таким образом, динамический диапазон входного речевого сигнала D_{in} будет составлять

$$D_{in} = 20 \cdot \log_{10} \frac{2^{15}}{1} \approx 90 \text{ дБ}. \qquad (4.1)$$

Для уменьшения динамического диапазона амплитуды кодируются в логарифмическом диапазоне от 0 до 60 дБ с равномерным шагом. Амплитуды в логарифмическом диапазоне вычисляются по следующему выражению:

$$A_{dB} = 20 \cdot \log_{10}(A_{Int}) - 30, \qquad (4.2)$$

где A_{dB} – амплитуды в логарифмическом диапазоне в децибелах; A_{Int} – целочисленные амплитуды.

Частоты определяются по спектру речевого сигнала и могут иметь следующие значения:

$$f = k \cdot \frac{F_S}{N_F}, \quad k = \overline{1, \frac{N_F}{2}}, \qquad (4.3)$$

где F_S – частота дискретизации; N_F – длина дискретного преобразования Фурье; k – индекс частотного отсчёта.

Следовательно, необходимо передавать из кодера в декодер только индекс частотного отсчёта k. Если длина дискретного преобразования Фурье N_F =1024, тогда значение k изменяется в диапазоне от 0 до 511 и должно быть представлено 9 битами. Для примера на рис. 4.4 отображены 7 частотных составляющих (условно изображены кружками), амплитуды и частоты которых пронормированы от 0 до 511.

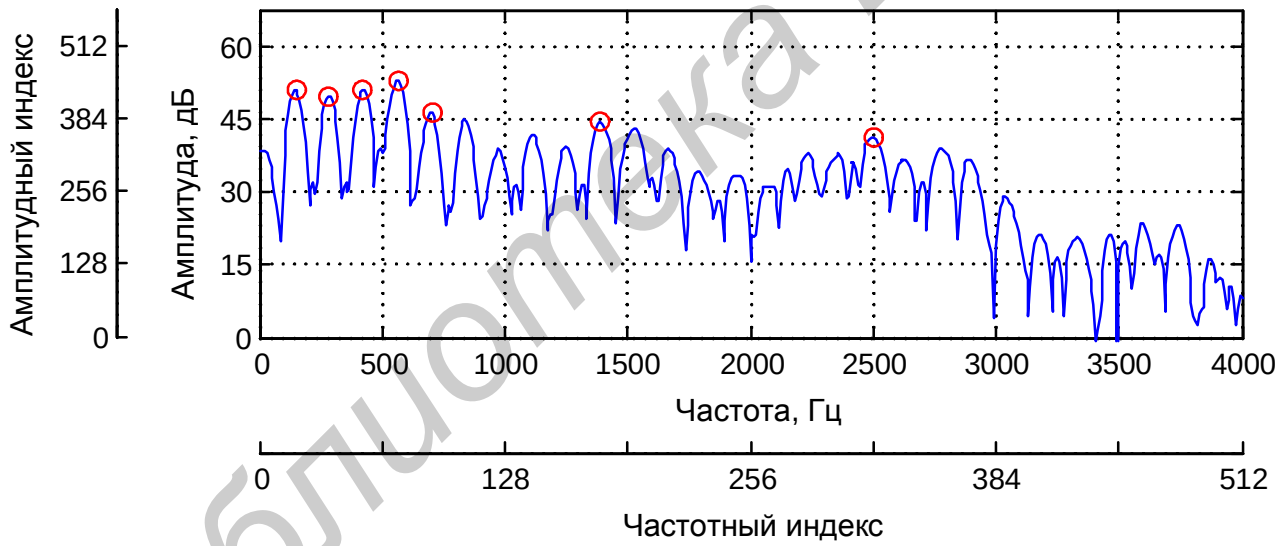

Рис. 4.4. Пример набора синусоидальных параметров для одного фрейма

Как уже было сказано ранее, исходя из свойств речевого сигнала, частота может быть закодирована только в диапазоне от 40 до 3800 Гц (частотный диапазон речи телефонного качества). Однако, поскольку применяется двухмерное векторное квантование, то это требует одинакового шага квантования по амплитуде и частоте. Таким образом, перед квантованием амплитуды должны быть пронормированы следующим образом:

$$A = A_{dB} \cdot \frac{k_{\max}}{A_{\max}} = A_{dB} \cdot \frac{N_F}{2 \cdot A_{\max}}, \qquad (4.4)$$

где $k_{\max}$ – максимально возможная величина индекса частотного отсчёта (например, если длина дискретного преобразования Фурье N_F =1024, тогда $k_{\max} = 512$); $A_{\max}$ – максимально возможная величина амплитуды, в данном случае $A_{\max} = 60$.

Кодовая книга представляет собой набор из L предопределённых выходных векторов y_i:

$$C = \{y_i\}, \quad i = \overline{1, L} \quad , \tag{4.5}$$

где

$$y_i = \{A_i, k_i\}. \tag{4.6}$$

Ошибка квантования обычно определяется как среднеквадратичная ошибка. В данном случае ошибка квантования $d(x, y)$ для входного вектора x и вектора из кодовой книги y определяется как квадрат разности координат (т.е. соответствующих значений амплитуды и частоты) этих векторов и вычисляется по следующей формуле:

$$d(x, y) = \frac{1}{2}\left(A^x - A^y\right)^2 + \frac{1}{2}\left(k^x - k^y\right)^2, \tag{4.7}$$

где A^x, A^y – нормированные амплитуды синусоид для входного вектора и ближайшего вектора из кодовой книги соответственно; k^x, k^y – индексы частотных отсчётов синусоид для входного вектора и ближайшего вектора из кодовой книги соответственно.

Для поиска оптимального элемента в кодовой книге используется критерий ближайшего соседа:

$$\begin{aligned} & q(x) = y_i \,, \\ & only \quad if \quad d(x, y_i) \le d\left(x, y_j\right), \\ & j \ne i, \quad 1 \le j \le L \,, \end{aligned} \tag{4.8}$$

где $q(\cdot)$ – оператор квантования.

Описанный выше подход для квантования синусоидальных параметров даёт хорошие результаты, если длина кодовой книги равняется 4096 или больше. Однако в этом случае данный алгоритм квантования параметров имеет высокую вычислительную сложность, что усложняет его использование в системах реального времени. Если длина кодовой книги очень велика, то возникают определенные трудности в процессе тренировки кодовой книги.

С другой стороны, если кодовая книга слишком мала, то выходная синтезированная речь имеет плохое качество, появляются значительные искажения. Кроме того, экспериментальным образом установлено, что ошибка по частоте намного сильнее влияет на качество синтезируемой речи, чем ошибка по ам-

плитуде. Вот почему в данном случае используется комбинация векторного квантования амплитуд и частот с дополнительным квантованием ошибки по частоте. Процесс коррекции частоты в этом случае можно представить, как показано на рис. 4.5.

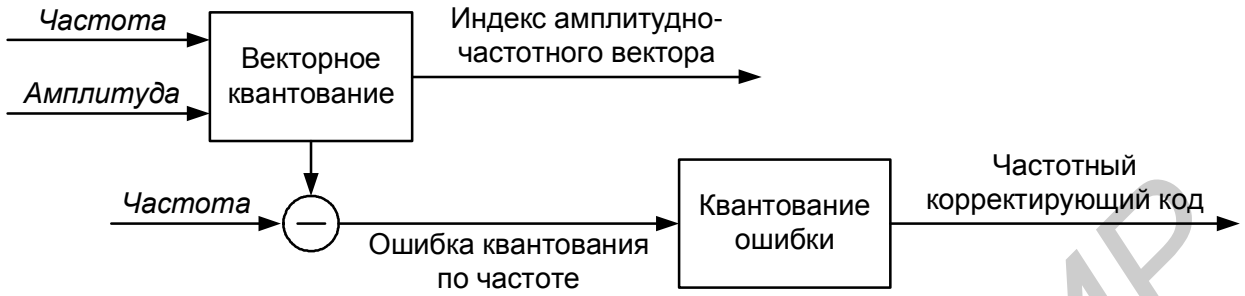

Рис. 4.5. Схема коррекции частоты

Фазы квантуются, используя скалярное квантование. Учитывая периодичность тригонометрических функций, фазы могут квантоваться только в диапазоне от $-\pi$ до π. Как показывают эксперименты, для получения хороших результатов при таком подходе для представления фаз достаточно 2-3 бита на одну составляющую.

4.3. Тренировка кодовой книги

Для тренировки кодовой книги используется описанный выше итерационный *алгоритм К-средних*. Целью применения этого алгоритма в данном случае является разделение набора обучающих векторов $\{x_n\}$, $n=\overline{1,P}$ на L кластеров R_i, где $i=\overline{1,L}$, таким образом, что удовлетворяются два необходимых условия оптимальности. Для этого сначала формируется исходная кодовая книга с набором начальных кодовых векторов $C(0)=\{y_i(0)\}$. Далее, с применением правила “ближайшего соседа”, выполняется классификация набора обучающих векторов $\{x_n\}$ по кластерам $R_i(m)$ и производится коррекция кодового вектора.

В данной процедуре тренировки кодовой книги два фактора играют очень важную роль: количество векторов для тренировки и число итераций в алгоритме. Оценка количества векторов для тренировки кодовой книги достаточно сложная задача, т.к. их число зависит главным образом от области применения данного алгоритма и индивидуально в каждом конкретном случае. Так, если используется относительно небольшое количество векторов для тренировки (меньше требуемого минимума), то не удастся аппроксимировать статистические характеристики последовательности входных векторов, т.е. при таких условиях нельзя получить хороший векторный квантователь. Хороший результат можно получить только при достаточно большом размере тренирующего набора векторов P, когда отношение величины P к числу элементов в кодовой книге L больше 50 и меньше 200. Что касается числа итераций для данного алгоритма, то необходимо иметь в виду, что перетренированная кодовая книга также не

будет давать хороших результатов и, таким образом, можно получить большую ошибку квантования, поскольку такая книга будет натренирована строго на определённый набор векторов и будет плохо представлять любой другой вектор, не входящий в этот набор.

Исходная кодовая книга может формироваться случайным образом на основе тренировочного множества или в соответствии с некоторым правилом. Например, можно всё амплитудно-частотное пространство жёстко разбить на определённое количество регионов (кластеров). Набор используемых значений частот и минимальный шаг по частоте определяются длиной применяемого при анализе речевого сигнала дискретного преобразования Фурье. Векторы для исходной кодовой книги формируются из значений амплитуд и частот в центрах данных кластеров.

Пример: дано 30 векторов для тренировки, необходимо натренировать 4-разрядную кодовую книгу, используя алгоритм *K*-средних. Ниже на рисунках отображены набор тренировочных векторов и центроиды (обозначены знаком *) на различных этапах тренировки. Границы кластеров показаны условно.

На начальном этапе формируется исходная (начальная) кодовая книга. Элементы кодовой книги выбираются из векторов тренировочного множества случайным образом. Затем все элементы тренировочного множества распределяются по кластерам по правилу “ближайшего соседа”. Как видно из рис. 4.6, тренировочные векторы распределяются по мере близости к центроидам (выбранные ранее элементы кодовой книги). Далее запускается итерационный алгоритм для оптимизации кодовой книги. Для этого в каждом цикле пересчитывается местоположение каждой центроиды и заново перераспределяются все тренировочные векторы. На рис. 4.7 отображены элементы кодовой книги (знаком *) и множество тренировочных векторов после первой итерации. Как видно из рис. 4.7, по сравнению с предыдущим состоянием изменились не только местоположения центроид, но и изменилось количество элементов в кластерах (произошло перераспределение – в каком-то кластере количество элементов увеличилось, а в каком-то уменьшилось). Таким образом, в процессе тренировки на каждой итерации пересчитываются местоположения центроид и перераспределяются элементы от кластера к кластеру (рис. 4.8) до тех пор, пока будет происходить уменьшение общего искажения. Когда вновь полученная ошибка квантования практически не будет отличаться от предыдущей, процесс оптимизации прерывается. Пример натренированной при помощи вышеописанного подхода кодовой книги с 256 элементами представлен на рис. 4.9. Элементы кодовой книги схематически изображены кружками. Первый вектор этой кодовой книги содержит нулевую амплитуду и частотный индекс k=1.

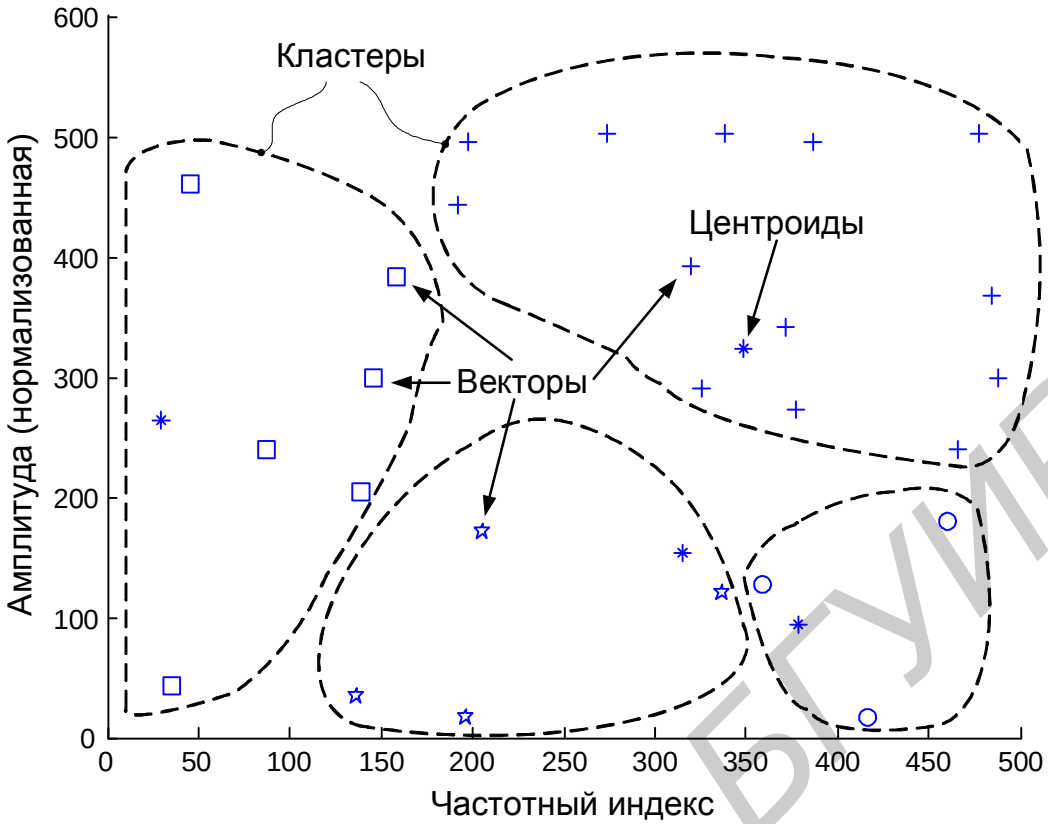

Рис. 4.6. Формирование исходной кодовой книги

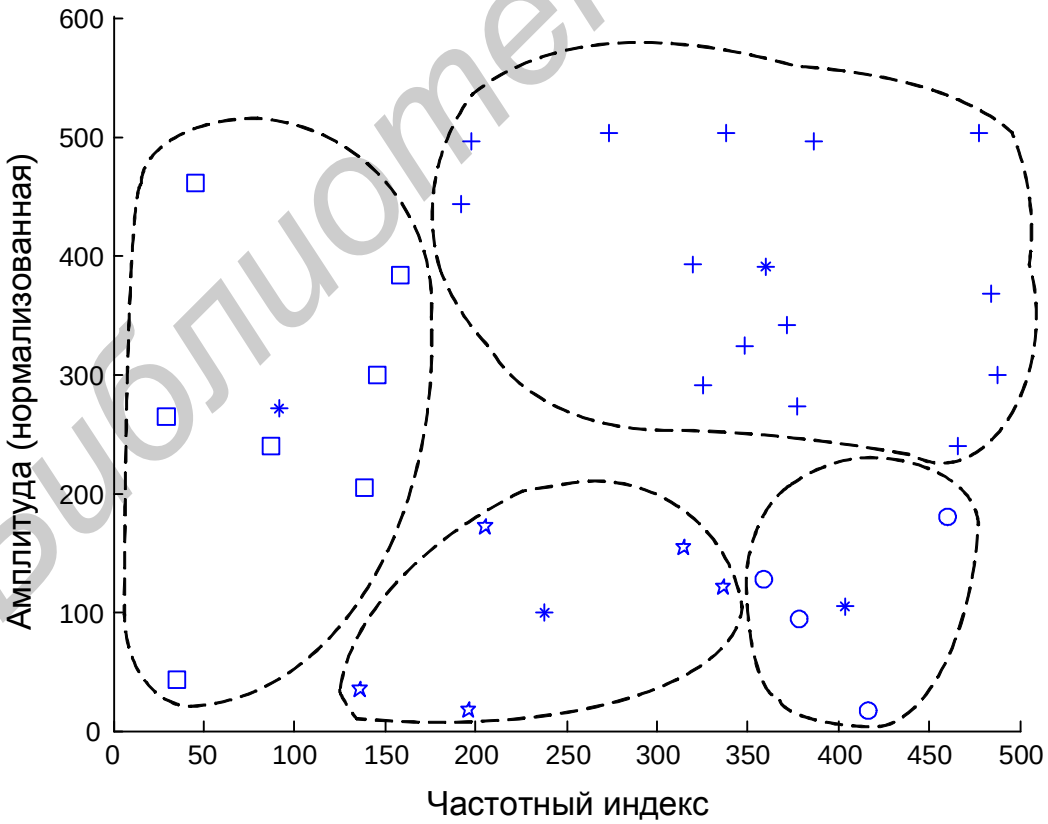

Рис. 4.7. Расположение кластеров после первой итерации

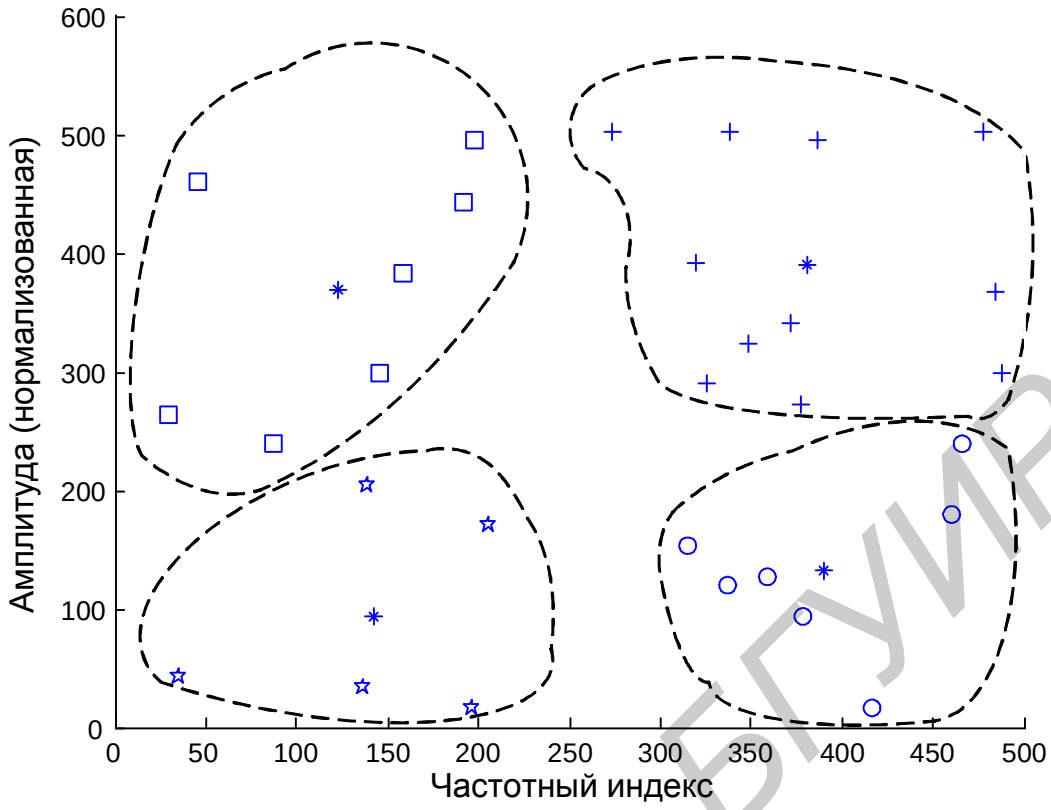

Рис. 4.8. Оптимальная кодовая книга

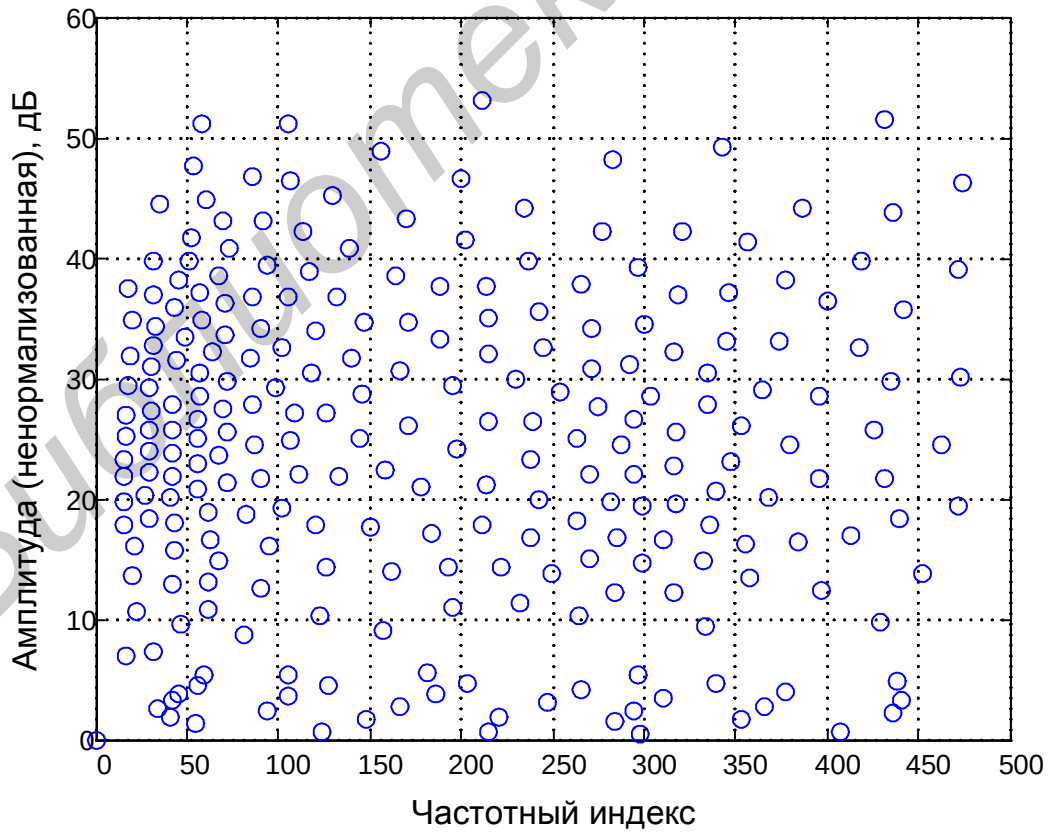

Рис. 4.9. Пример кодовой книги с 265 элементами

5. ОБЪЕКТИВНАЯ ОЦЕНКА КАЧЕСТВА РЕЧЕВЫХ КОДЕРОВ

5.1. Схема объективной оценки качества реконструированного сигнала кодера

Методы и алгоритмы формирования объективных оценок качества речевого сигнала должны быть инвариантны к типу кодера, а результаты их работы коррелировать с субъективными оценками качества восстановленного сигнала кодера. Тесты для проверки качества аудиокодера регламентируются по *ITU-R Recommendation BS.1116, 1997*. В настоящее время используется новый стандарт для обеспечения оценки качества перцептуального аудиокодера *PEAQ* – Perceptual Evaluation of Audio Quality, схема которого базируется на периферийной модели уха человека (рис. 5.1).

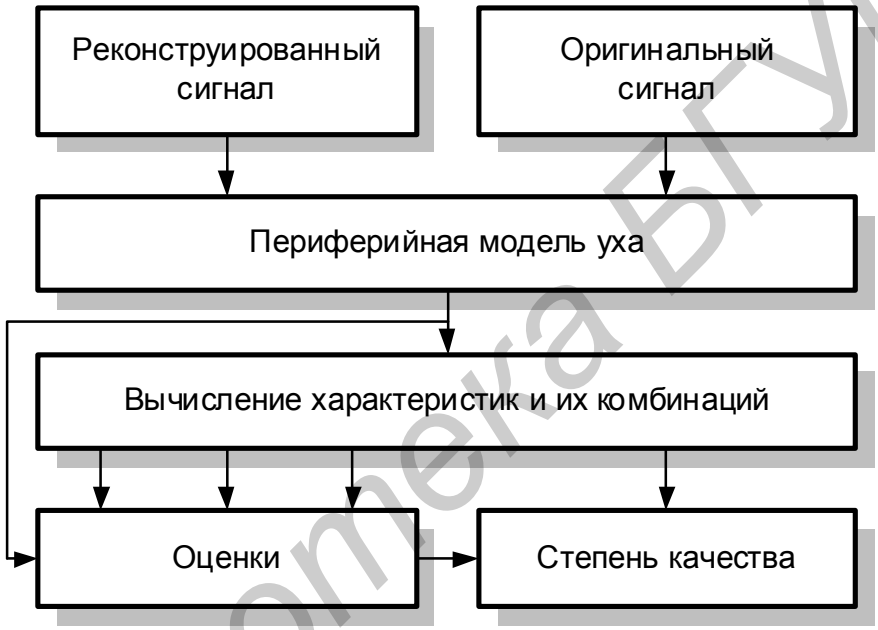

Рис. 5.1. Схема объективной оценки кодера

Таким образом, обработка сигнала ведется в критических частотных полосах (барках) E_{CB}, формирование которых осуществляется с помощью пакета дискретного вэйвлет-преобразования (ПДВП) или на основании сжатого дискретного преобразования Фурье (*WDFT* – Warped Discrete Fourier Transform). Анализ входных сигналов в критических частотных полосах выполняется на основе ПДВП и *WDFT*. Вся полоса сигнала 4000 Гц разделяется на 18 частотных полос, или 18 барков. Перцептуальное качество моделей, реализованных на ПДВП и *WDFT*, выше, чем у известной модели Джонсона, базирующейся на БПФ.

Соответствие спектральных компонент для *WDFT* и *FFT* для речевого сигнала с частотой дискретизации 8 кГц и длине преобразования 256 отсчетов показано в таблице.

Соответствие спектральных компонент для *WDFT* и *FFT*

Критические полосы	Спектраль-ные линии	Кол-во	Частоты, Гц	Спектраль-ные линии	Кол-во	Частоты, Гц
1	1 – 7	7	13 – 92	1 – 3	3	31 – 94
2	8 – 15	8	105 – 198	4 – 6	3	125 – 188
3	16 – 22	7	212 – 294	7 – 9	3	219 – 281
4	23 – 29	7	308 – 394	10 – 12	3	313 – 375
5	30 – 36	7	408 – 498	13 – 16	4	406 – 500
6	37 – 44	8	514 – 627	17 – 20	4	531 – 625
7	45 – 52	8	644 – 768	21 – 24	4	656 – 750
8	53 – 59	7	786 – 904	25 – 29	5	781 – 906
9	60 – 67	8	925 – 1080	30 – 34	5	938 – 1063
10	68 – 74	7	1103 – 1255	35 – 40	6	1094 - 1250
11	75 – 81	7	1282 – 1458	41 – 47	7	1281 - 1469
12	82 – 88	7	1489 – 1694	48 – 55	8	1500 - 1719
13	89 – 95	7	1731 – 1972	56 – 64	9	1750 - 2000
14	96 – 102	7	2015 – 2301	65 – 74	10	2031 - 2313
15	103 – 109	7	2352 – 2688	75 – 86	12	2344 - 2688
16	110 – 116	7	2748 – 3134	87 – 100	14	2719 - 3125
17	117 – 123	7	3202 – 3629	101 – 118	18	3156 - 3688
18	124 – 128	5	3703 – 4000	119 – 128	10	3719 - 4000

На рис. 5.2 показано дерево ПДВП, осуществляющего разделение частотного интервала сигнала на полосы согласно критической шкале частот:

$$CB\text{-}WPD\text{: } (l,n) \in E_{CB},\ l = 0\ldots6, \qquad (5.1)$$

где E_{CB} – обозначает множество узлов дерева ПДВП соответствующего *CB-WPD*. Дерево *CB-WPD* делит частотный диапазон [0 – 4 кГц] на 18 неравномерных полос *CBW*(*f*), т.е. на 18 барков. Корневой узел (*l*,*n*) = (0,0) данного дерева соответствует всему частотному диапазону сигнала. Каждый внутренний узел дерева $(l,n) \in E_{CB}$, названный узлом предка, делится на два потомка: 1-й потомок и 2-й потомок, ассоциируемые соответственно с высокочастотной и низкочастотной фильтрацией, выходные сигналы (вэйвлет-коэффициенты) которых децимируются в соотношении 2.

5.2. Объективные оценки качества сигнала

5.2.1. Соотношение сигнал – шум (SNR)

Объективные оценки качества алгоритмов кодирования сигнала часто базируются на соотношениях: сигнал/шум (*SNR*) (рис. 5.3) и сегментный *SEGSNR*.

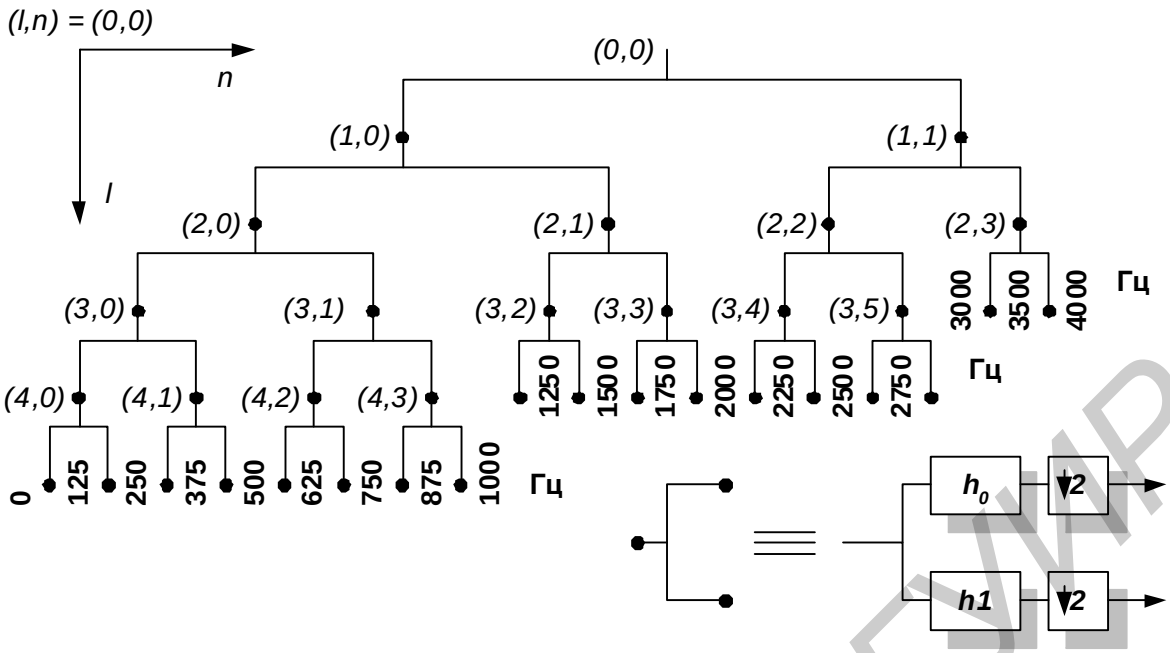

Рис. 5.2. ПДВП $(l,n)\in E_{CB}$

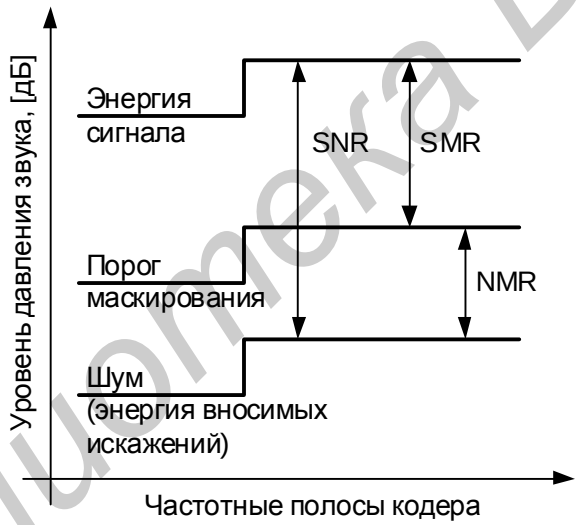

Рис. 5.3. Объективные оценки качества сигнала

Значение *SNR* для каждой полосы критических частот j=1…K рассчитывается как

$$SNR(j)=\frac{X^2(j)}{s^2(j)}. \tag{5.2}$$

Для всего анализируемого сигнала значение *SNR* рассчитывается как среднее арифметическое

$$SNR=\frac{1}{N}\sum_{i=0}^{N}\frac{1}{K}\sum_{j=0}^{K}SNR^{(i)}(j), \tag{5.3}$$

где K – число критических полос, N – число фреймов.

Более адекватной оценкой *segmental SNR* (*SEGSNR*), которая вычисляется как среднее геометрическое значение энергии *SNR* на коротком участке.

$$SEGSNR = \left(\prod_{i=0}^{N} \frac{1}{K} \sum_{j=0}^{K} SNR^{(i)}(j) \right)^{\frac{1}{N}}. \qquad (5.4)$$

Однако данный способ измерения становится незначимым для новой генерации алгоритмов кодирования, выполняющих компрессию сигнала в пределах от 2 до 8 Кбит/с, так как восстановленный не похож на оригинальный сигнал. Кодерами данного типа преследуется цель адекватного воспроизведения перцептуально значимых аспектов сигнала с сохранением разборчивости и натуральности.

Оценка качества сигнала *SNR* не отражает объективной картины качества восстановленного сигнала кодером речи. Эксперименты с тональными сигналами доказывают несостоятельность данной оценки качества.

5.2.2. Соотношение шум – порог маскирования (NMR)

Большинство оценок качества восстановленного сигнала базируется на отношении энергии шума (энергии вносимых искажений) к порогу маскирования *NMR* (см. рис. 5.3). Данное отношение является оценкой расстояния между фактическими искажениями и максимально неслышимыми искажениями. Исследования последних лет показывают высокую коррелированность *NMR* с субъективными тестами.

Оценка *NMR* *i*-го фрейма сигнала $x_i(n/f_S)$ вычисляется по следующей схеме (рис. 5.4).

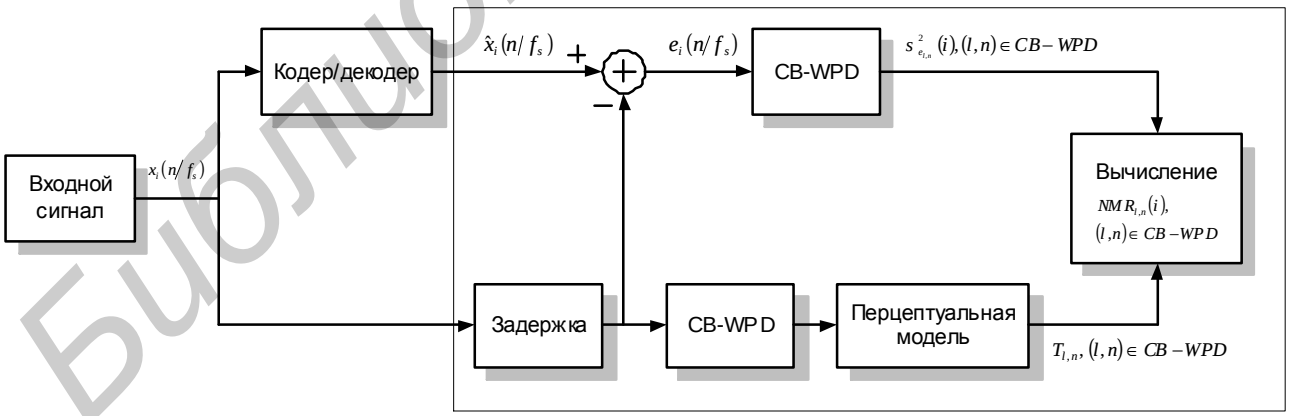

Рис. 5.4. Блок-схема вычисления $NMR_{l,n}(i)$, $(l,n)\in E_{CB}$

Определяются ошибка восстановленного сигнала $e_i(n/f_S)=\hat{x}_i(n/f_S)-x_i(n/f_S)$ и её дисперсия в каждой критической частотной полосе $s^2_{e_{l,n}}(i)$, $(l,n)\in E_{CB}$. По параллельному каналу для оригинального аудиосигнала вычисляются пороги

маскирования $T_{l,n}(i)$, $(l,n) \in E_{CB}$ в соответствующих критических полосах на основании процедуры расчета порогов маскирования в вэйвлет-области. На базе данных оценок для *i*-го фрейма находится соотношение

$$NMR_{l,n}(i) = \frac{s^{2}_{e_{l,n}}(i)}{T_{l,n}(i)}, \quad (l,n) \in E_{CB} \tag{5.5}$$

и среднее арифметическое по всем критическим частотным полосам для *i*-го фрейма:

$$NMR_{loc}(i) = 10 \cdot \log_{10}\left(\frac{1}{K} \cdot \sum_{\substack{\text{для} \\ \forall (l,n) \in CB-WPD}} NMR_{l,n}(i) \right), \text{ дБ.} \tag{5.6}$$

Объективные показатели, характеризующие качество кодирования речи кодером для полного сигнала, определяются как среднее арифметическое $NMR_{loc}(i)$:

$$NMR_{total} = 10 \cdot \log_{10}\left(\frac{1}{N} \sum_{i=1}^{N} 10^{(NMR_{loc}(i)/10)} \right), \text{ дБ} \tag{5.7}$$

или как среднее геометрическое $NMR_{loc}(i)$:

$$NMR_{SEG} = \frac{1}{N} \sum_{i}^{N} NMR_{loc}(i), \text{ дБ.} \tag{5.8}$$

Негативная величина оценок NMR_{total} или NMR_{SEG} показывает оценку нижней границы порога восприятия, а позитивная величина данных значений является оценкой энергии воспринимаемых искажений.

5.3. Перцептуальные оценки искажений спектра барков

5.3.1. Оценка искажений спектра барков

Перцептуальная оценка искажений спектра барков (*BSD* – Bark spectral distortion) представляет собой усредненную величину воспринимаемых человеком искажений, присутствующих в восстановленном речевом сигнале. Слух человека обладает неодинаковой чувствительностью к энергии на разных частотах. Графически этот факт можно представить в виде кривых равной громкости, которые показаны на рис. 5.5. Вдоль каждой кривой уровень громкости, измеряемый в фонах, остается постоянным и полагается равным уровню звукового давления в децибелах на частоте 1 кГц.

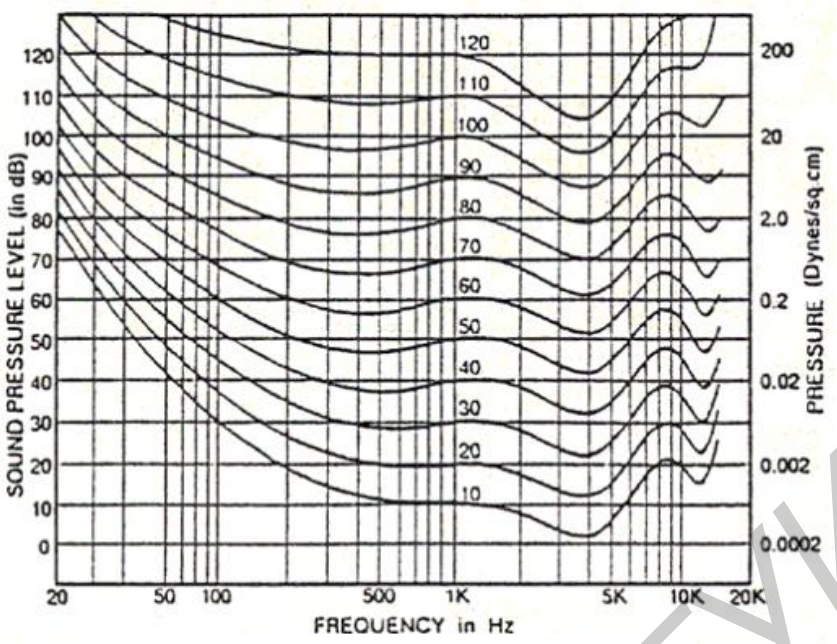

Рис. 5.5. Кривые равной громкости

Как видно, кривые семейства схожи между собой, и их можно аппроксимировать абсолютным порогом слышимости, поднимая его по амплитуде. Исходя из этого компенсация частотной зависимости чувствительности к энергии (преобразование децибелы – фоны) выполнялась в виде

$$P = D - ATH + ATH_{1\,kHz}, \tag{5.9}$$

где D и P – амплитуды спектральной компоненты в децибелах и фонах соответственно; ATH и $ATH_{1\,kHz}$ – значения абсолютного порога слышимости на частотах данной спектральной компоненты и 1 кГц в децибелах соответственно:

$$ATH(f) = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 10^{-3} f^4, \tag{5.10}$$

где f – частота в килогерцах.

Схема вычисления *BSD* показана на рис. 5.6.

Процесс вычисления *BSD* по данной схеме состоит из двух основных этапов: вычисления уровней громкости оригинального $x(n)$ и восстановленного сигналов $y(n)$ и блока вычисления оценки *BSD*. В соответствии с рис. 5.7 входной сигнал (восстановленный сигнал) анализируется в критических частотных полосах банками фильтров, реализованными на преобразованиях *ПДВП* или *WDFT*.

Мощность коэффициентов $X(z)$ преобразования рассчитывается как

$$D(z) = \left| X(z) \right|^2. \tag{5.11}$$

В порядке перцептуальной обработки сигнала речи кривые уровней равной громкости в децибелах должны быть конвертированы в фоны.

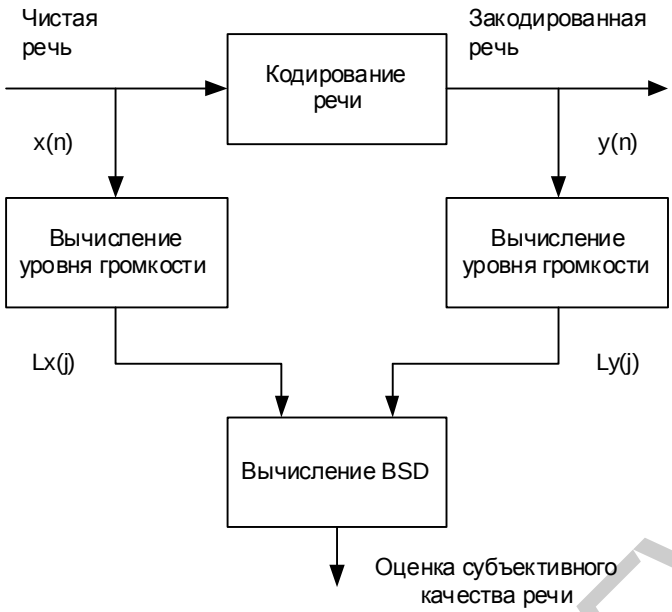

Рис. 5.6. Схема вычисления оценки качества речи *BSD*

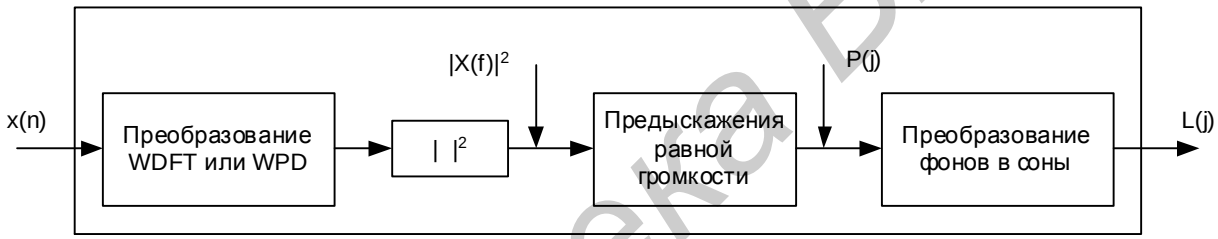

Рис. 5.7. Схема преобразования входного сигнала в субъективные уровни громкости в сонах

Для речи, передаваемой по телефонной линии, частотная полоса ограничивается диапазоном 300…3400 Гц, а уровни интенсивности – 40…80 дБ. В среднем сочетание полос представляет собой прямоугольный блок на уровнях равной громкости в районе 1800 Гц. Предыскажения равной громкости вычисляются от энергии входного сигнала $D(z)$. Уровни интенсивности $D(z)$ трансформируются в область фонов в интересующем интервале частот. Билинейный фильтр предыскажений, применяемый для выделения уровней громкости в фонах, имеет следующую передаточную характеристику:

$$H(z)=\frac{2{,}6+z^{-1}}{1{,}6+z^{-1}}. \qquad (5.12)$$

Полученное значение взвешенной энергии спектра $P(z)=H(z)\cdot D(z)$ выравнивается абсолютным порогом слышимости (5.10) для выражения энергии спектра в уровнях громкости, которая описывается шкалой фонов. Отношение субъективной громкости в сонах на основании уровней громкости в фонах выражается как

$$L = \begin{cases} 2^{(P-40)/10}, & \text{если} \quad P \geq 40; \\ \left(P/40\right)^{2.642}, & \text{если} \quad P < 40. \end{cases} \tag{5.13}$$

Графическая зависимость субъективной громкости в сонах от уровней громкости в фонах показана на рис. 5.8.

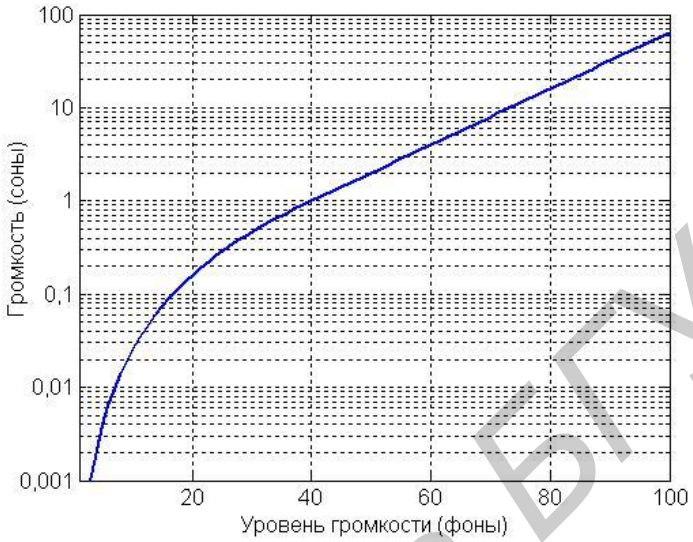

Рис. 5.8. Зависимости громкости в сонах от уровня громкости в фонах

Оценка *BSD* определяет искажения как квадрат евклидова расстояния между оцененными субъективными громкостями оригинального $L_x^{(i)}(j)$ и кодированного сигналов $L_y^{(i)}(j)$ для каждой критической частотной полосы. Общая оценка *BSD* для всего анализируемого сигнала задается выражением

$$BSD = \frac{\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{K}\left[L_x^{(i)}(j) - L_y^{(i)}(j)\right]^2}{\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{K}\left[L_x^{(i)}(j)\right]^2}, \tag{5.14}$$

где N – количество фреймов; K – число частотных полос; $L_x^{(i)}(j)$ – барк-спектр i-го фрейма оригинального сигнала; $L_y^{(i)}(j)$ – барк-спектр i-го фрейма закодированного сигнала.

5.3.2. Модифицированная оценка искажений спектра барков

Перцептуально-модифицированная оценка искажений спектра барков (*MBSD* – Modify Bark Spectral Distortion) использует перцептуальную модель слуха человека для определения слышимых искажений, тогда как *BSD* манипулирует эмпирически определенной оценкой энергетического порога. Другое существенное отличие *MBSD* от *BSD* состоит в том, что искажения вычисляют-

ся как средняя разность субъективных оценок громкости. Перцептуальная метрика искажений не определяется при рассмотрении оценки *BSD*. Алгоритм реализации вычисления *MBSD* показан на рис. 5.9.

Порог шума рассчитывается в блоке психоакустической информации и сравнивается с разностью субъективных оценок громкости оригинального $L_x^{(i)}(j)$ и восстановленного $L_y^{(i)}(j)$ для определения, воспринимаются ли данные искажения на слух. Когда разность субъективных оценок громкости меньше порога шума, то искажения не слышимы, в противном же случае слышимы.

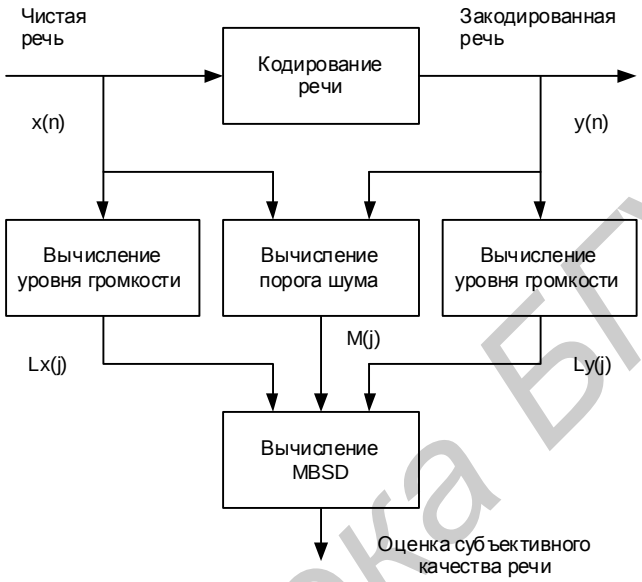

Рис. 5.9. Алгоритм вычисления оценки качества речи *MBSD*

Значение оценки *MBSD* определяется следующим выражением:

$$MBSD = \frac{1}{N_x}\sum_{i=0}^{N}\left[\sum_{j=0}^{K} M(j)\left|L_x^{(i)}(j) - L_y^{(i)}(j)\right|^2\right], \qquad (5.15)$$

где N – количество фреймов; K – число полос; $M(j)$ – показатель присутствия искажений в j-й полосе; $L_x^{(i)}(j)$ – барк-спектр i-го фрейма оригинального сигнала; $L_y^{(i)}(j)$ – барк-спектр i-го фрейма закодированного сигнала. Параметр $M(j) = 0$, когда искажения в полосе j не воспринимаются на слух, $M(j) = 1$ в противном случае.

Корреляционная зависимость субъективных оценок с объективными оценками качества сигнала облегчают процесс предсказания *MOS* (Mean Opinion Score) – оценки мнений экспертов, которые осуществляют прослушивание оригинальных и восстановленных сигналов. Градация оценок, полученных способом *MOS*, выражается, как правило, числами от 5 до 1 и ассоциируется с численными (объективными) значениями, описывающими уровень искажений.

6. РЕКОМЕНДУЕМЫЕ ЛАБОРАТОРНЫЕ РАБОТЫ

Лабораторная работа №1. Исследование методов детектирования речи

Цель работы: исследовать характеристики работы детектора речи в условиях аддитивного шума с различными *SNR*.

Порядок выполнения работы

1. Реализовать в среде Matlab функцию, рассчитывающую объективную оценку работы детектора.

2. Исследовать зависимость характеристик работы детектора речи от соотношения сигнал-шум исходного речевого сигнала. В качестве тестовых сигналов берутся речевые сигналы с SNR 50, 20, 10, 5, 0, -5 дБ.

3. Сравнить характеристики работы одного из стандартных методов с характеристиками антропоморфного метода детектирования речи. В качестве стандартного метода взять один из следующих: 1) энергетический VAD; 2) спектральный VAD; 3) кепстральный интегральный VAD; 4) кепстральный дифференциальный VAD.

4. Исследовать влияние изменения параметров антропоморфического детектора речи на его работу. Необходимо исследовать влияние изменения параметров степени сжатия g и доверительной границы l на характеристики работы детектора. Изменение параметров производить в пределах [0,5, 5].

Лабораторная работа №2. Векторное квантование LSF-коэффициентов

Цель работы: исследовать методы векторного квантования для кодирования *LSF*-коэффициентов.

Порядок выполнения работы

1. Создать обучающее и тестовое множества (базы данных) параметров *LSF*. Исходные данные: речевые файлы в формате .WAV. На выходе должны быть 2 тренировочные базы данных *LSF*, полученные из мужских и женских голосов, и 2 тестовые базы данных. Количество векторов *LSF* в обучающих множествах определить исходя из тренировочного отношения для разрядности 6. Построить график изменения *LSF* от сегмента к сегменту внутри одного из файлов.

2. Рассчитать коэффициенты внутрисегментной корреляции по всей базе данных.

3. Провести тренировку кодовых книг по методу *SVQ*. Исходные данные: тренировочные базы данных. На выходе должны быть 2 структурированные кодовые книги разрядностью 6 и 2 кодовые книги – меньшей разрядности.

4. Рассчитать объем памяти, занимаемой кодовой книгой, в случае использования трех 6-разрядных кодовых субкниг (*SVQ*) и при использовании неструктурированной 18-разрядной кодовой книги.

5. Провести квантование векторов *LSF* в тестовых базах данных. Исполь-

зовать все варианты квантования (тестовое множество мужской речи по кодовой книге из мужской речи, мужской по женской и т.д.).

6. Оценка качества квантования во всех вариантах квантования (использовать логарифмическое искажение спектра).

7. Оценить время поиска в кодовых книгах разной размерности. Проанализировать полученные результаты.

8. Для любого вектора *LSF* из тестовой базы данных получить АЧХ фильтра. Сравнить с АЧХ, полученной из квантованного вектора *LSF*. Провести квантование с использованием эталонной кодовой книги. Получить АЧХ.

Лабораторная работа №3. Векторное квантования параметров линейного предсказания в CELP-кодерах

Цель работы: оценить возможность применения векторного квантования для кодирования параметров линейного предсказания.

Порядок выполнения работы

1. С помощью прилагаемой программы *CELP*-вокодера сформировать обучающую и тестовую базу данных по речевым файлам, содержащим мужские и женские голоса дикторов.

2. По обучающему множеству построить векторный квантователь *LSF* в соответствии с вариантом задания:

Вариант	1	2	3	4	5	6
Порядок фильтра-предсказателя	8	10	12	14	16	18
Параметры расщепления	нет	4; 6	4; 4; 4	4; 4; 6	2; 2; 12	4; 4; 10
Глубина кодовых книг, векторов	16 32 64 128	16; 16 32; 16 64; 32 128; 64	16; 16; 16 32; 16; 16 64; 32; 16 128; 32; 32	16; 16; 16 32; 16; 32 32; 32; 32 64; 32; 32	32; 32; 32 64; 32; 32 128; 64; 64 128; 128; 64	64; 32; 32 64; 64; 64 128; 64; 64 128; 128; 64

3. Заквантовать обучающую и тестовую базу данных *LSF*.

4. Оценить качество квантования *LSF* по тренировочной и тестовой базе данных.

5. Загрузить в программу *CELP*-кодера квантованную версию тестовой базы данных, сгенерировать выходной файл с реконструированной речью.

6. Оценить качество реконструированной речи согласно методам объективных оценок качества речевого сигнала (см. разд. 5) с помощью специального программного обеспечения.

7. Провести исследование влияния параметров векторного квантования *LSF* на качество реконструированной речи в составе *CELP*-кодера в соответствии с вариантом задания, проанализировать полученные результаты.

Лабораторная работа №4. Исследование векторного квантования в синусоидальных кодерах речевого сигнала

Цель работы: исследовать применение векторного квантования параметров в синусоидальных кодерах речи.

Порядок выполнения работы

1. С помощью команды **load** загрузить в рабочее пространство Matlab три массива синусоидальных параметров: амплитуды, частоты и фазы (файлы **am.mat**, **fr.mat** и **ph.mat** соответственно).

2. Реализовать и исследовать скалярный квантователь для кодирования фаз. Получить квантованные значения фаз.

3. Реализовать и исследовать векторный квантователь без частотной коррекции, используя данную базу параметров (амплитуд и частот). Для этого сгенерировать исходную кодовую книгу (случайным образом) и провести оптимизацию кодовой книги по алгоритму *K*-средних. Получить квантованные значения амплитуды и частоты.

4. Реализовать и исследовать векторный квантователь с частотной коррекцей, используя для коррекции кодовую книгу длиной *N* элементов. Для этого сгенерировать исходную кодовую книгу (случайным образом) и провести оптимизацию кодовой книги по алгоритму *K*-средних. Получить квантованные значения амплитуды и частоты.

5. С помощью функции **decoder.m** осуществить синтез речевого сигнала, используя исходные и заквантованные параметры. Сравнить субъективное качество речи (на слух) для обоих случаев.

6. Оценить качество реконструированной речи согласно методам объективных оценок качества речевого сигнала (см. разд. 5) с помощью специального программного обеспечения.

Литература

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

2. Kondoz A.M. Digital speech: coding for low bit rate communication systems. – NY.: John Wiley & Sons, Inc., 1996. – 442 p.

3. Kleijn W.B., Palival K.K., eds. Speech coding and synthesis. – Amsterdam: Elsevier, 1995.

4. Zwicker E., Fastl H. Psychoacoustics: Fact and Models. – Berlin, Germany: Springer-Verlag, 1990. – 380p.

5. Rabiner L.R., Schafer R., Digital Processing of Speech Signals. – Englewood Cliffs, New Jersey: Prentice-Hall, 1979.

6. Gersho A., Gray R.M. Vector Quantization and Signal Compression – Boston: Kluwer Academic Press, 1992.

7. Витязев В.В. Цифровая частотная селекция сигналов. – М.: Радио и связь, 1993. – 240 с.

Св. план 2005, поз. 23

Учебное издание

Петровский Александр Александрович,
Лихачёв Денис Сергеевич,
Петровский Алексей Александрович и др.

Речевые интерфейсы ЭВС

Учебно-методическое пособие
для студентов специальности I – 40 02 02
«Электронные вычислительные средства»
дневной формы обучения

Редактор Е.Н. Батурчик

Подписано в печать 20.09.2005.	Формат 60х84 1/16.	Бумага офсетная.
Гарнитура «Таймс».	Печать ризографическая.	Усл. печ. л. 3,14.
Уч.-изд. л. 3,0.	Тираж 150 экз.	Заказ 160.

Издатель и полиграфическое исполнение: Учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники»
Лицензия на осуществление издательской деятельности №02330/0056964 от 01.04.2004.
Лицензия на осуществление полиграфической деятельности №02330/0131518 от 30.04.2004.
220013, Минск, П. Бровки, 6