ИТМО

М. Б. Столбов, В. Л. Иванов

АНАЛИЗ И МОДЕЛИ РЕЧЕВЫХ СИГНАЛОВ

Санкт-Петербург
2024

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

М. Б. Столбов, В. Л. Иванов

АНАЛИЗ И МОДЕЛИ РЕЧЕВЫХ СИГНАЛОВ

УЧЕБНОЕ ПОСОБИЕ

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлению подготовки 09.04.02
«Информационные системы и технологии»
в качестве учебно-методического пособия
для реализации основных профессиональных образовательных программ
высшего образования магистратуры

Санкт-Петербург
2024

Столбов М. Б., Иванов В. Л., Анализ и модели речевых сигналов – СПб.: НИУ ИТМО, 2024. – 97 с.

Анализ и модели речевых сигналов

Рецензент:

Рыбин Сергей Витальевич, кандидат физико-математических наук, доцент (квалификационная категория «доцент практики») факультета информационных технологий и программирования, Университета ИТМО.

АННОТАЦИЯ

В книге изложены материалы второй части курса лекций «Цифровая обработка речевых сигналов», прочитанных в течение ряда лет студентам, обучающимся по направлению «Информационные системы и технологии».

Книга предполагает знакомство с курсом «Цифровая обработка сигналов».

В книге приведены основные термины и понятия, а также даны рекомендации по анализу речевых и акустических сигналов.

Материал книги предназначен прежде всего тем, кто предполагает работать в области речевых технологий – решении задач автоматического распознавания речи, синтеза речи, идентификации дикторов. Книга также будет полезна тем, кто занимается анализом и обработкой акустических сигналов и анализом акустических событий.

ИТМО (Санкт-Петербург) — национальный исследовательский университет, научно-образовательная корпорация. Альма-матер победителей международных соревнований по программированию. Приоритетные направления: IT и искусственный интеллект, фотоника, робототехника, квантовые коммуникации, трансляционная медицина, Life Sciences, Art&Science, Science Communication.

Лидер федеральной программы «Приоритет-2030», в рамках которой реализовывается программа «Университет открытого кода». С 2022 ИТМО работает в рамках новой модели развития — научно-образовательной корпорации. В ее основе академическая свобода, поддержка начинаний студентов и сотрудников, распределенная система управления, приверженность открытому коду, бизнес-подходы к организации работы. Образование в университете основано на выборе индивидуальной траектории для каждого студента.

ИТМО пять лет подряд — в сотне лучших в области Automation & Control (кибернетика) Шанхайского рейтинга. По версии SuperJob занимает первое место в Петербурге и второе в России по уровню зарплат выпускников в сфере IT. Университет в топе международных рейтингов среди российских вузов. Входит в топ-5 российских университетов по качеству приема на бюджетные места. Рекордсмен по поступлению олимпиадников в Петербурге. С 2019 года ИТМО самостоятельно присуждает ученые степени кандидата и доктора наук.

ОГЛАВЛЕНИЕ

Предисловие

Речевые технологии интенсивно развиваются. Новые направления делают актуальной задачу освоения новых методов в учебном процессе. Данная книга представляет собой такую попытку сделать шаг в данном направлении.

Книга написана в первую очередь для магистрантов, обучающихся по образовательной программе «Речевые технологии и машинное обучение».

Пособие должно помочь в освоении материала 2-й части курса лекций «Цифровая обработка речевых сигналов». В книге изложены термины и понятия кратковременного анализа акустических и речевых сигналов. Эта книга является фактически продолжением материала, изложенного в книге «Основы анализа и обработки речевых сигналов» (Столбов М. Б.– СПб.: НИУ ИТМО, 2021), применительно к области анализа и обработки нестационарных акустических и речевых сигналов.

Материал данной книги является основой как для освоения последующих частей курса цифровой обработки речевых сигналов (обработка акустических и речевых сигналов, компенсация искажений речевых сигналов, микрофонные решетки), так и для следующих дисциплин специальности (автоматическое распознавание речи, распознавание дикторов, распознавание акустических событий и других).

Материал книги выходит за рамки анализа исключительно речевых сигналов, и будет полезен для приобретения необходимых знаний и практической работы в области биометрических технологий, фоноскопии, решения задач обнаружения акустических событий, анализа акустических сцен, неразрушающего акустического контроля. Книга также может быть полезной для слушателей курсов повышения квалификации и переподготовки специалистов-практиков, работающих в указанных областях.

Параллельно с теоретическим материалом обучающимся предлагались лабораторные работы, целью выполнения которых было освоение практических навыков оценки характеристик сигналов и представления полученных результатов в понятном виде, соотнесение их с теоретическими представлениями и умение делать из них обоснованные выводы. Описание лабораторных работ содержится в книге «Цифровая обработка речевых сигналов: Учебно-методическое пособие по лабораторному практикуму» (Столбов М. Б., Кассу А.-Р. М – СПб.: НИУ ИТМО, 2016).

Одной из особенностей книги является то, что помимо характеристик непосредственно речевых сигналов мы рассматриваем характеристики, относящиеся к основным этапам формирования цифрового речевого сигнала: генерации, распространению, аналоговой обработке и звукозаписи.

В предыдущие десятилетия много внимания уделялось написанию алгоритмов. Теперь же исследователю доступна большая часть алгоритмов. Однако важно понимание назначения и смысла этих алгоритмов. В книге мы пытались уделить внимание этому аспекту.

В литературе по анализу речевых сигналов существует тенденция уделять большое внимание алгоритмам вычислений. Однако сейчас почти никто не пишет собственные коды, предпочитая использовать готовые процедуры. В книге не приводятся детали вычислений, поскольку современные библиотеки (фреймворки) включают в себя сотни готовых вычислительных методов. Численные методы обсу-

ждаются лишь настолько, чтобы читатель понял, в чем смысл вычислений.

Центральным является понятие кратковременного анализа сигналов, являющегося основой большинства методов обработки речевых сигналов.

Содержание книги является теоретическим обучающим материалом для лабораторного практикума, то есть практико-ориентированным. Мы руководствовались тезисом «От теории к практике». Каждая глава сопровождается задачами и упражнениями.

Помимо прочего подробно рассмотрены понятия кепстра и модуляционного спектра, недостаточно освещенные в русскоязычной литературе.

В главе 1 излагаются необходимые теоретические основы кратковременного анализа – основы анализа нестационарных сигналов. Рассмотрены методы кратковременной оценки базовые параметров сигналов – энергии, мощности, огибающиих и ряд других.

Глава 2 посвящена центральной процедуре кратковременного анализа сигналов – кратковременному спектральному анализу. Отдельный раздел посвящен модуляционному спектру и методам его оценки.

В главе 3 рассмотрен кепстр – эффективный инструмент анализа акустических сигналов, недостаточно освещенный в литературе. Основное внимание уделено рассмотрению физического смысла кепстра и областям его практического применения.

В главе 4 описаны методы оценки параметров и модели речевых сигналов.

В главе 5 дан обзор методов оценки качества и разборчивости речевых сигналов, применяемых в различных областях речевых технологий, в частности, в задачах анализа и синтеза речевых сигналов.

В заключение авторам хотелось бы выразить благодарность Василию Коняхину за ценные советы и помощь в оформлении рукописи к изданию.

Глава 1
Кратковременный анализ речевых сигналов

Целью анализа сигналов является численная оценка их свойств, описываемых различными параметрами и характеристиками. Средние значения параметров характеризуют постоянство, локальные – изменчивость во времени.

Рассматриваемые в данной главе характеристики являются локальными или «мгновенными», то есть относятся к определенному моменту времени, а не ко всему сигналу. А сам анализ называется «кратковременным» *(short-time analysis*).

Кратковременный анализ речевых сигналов может быть применен к любым характеристикам сигналов, прежде всего к параметрам их функций распределения:

– среднее значение (*mean*);
– максимум и минимум (*maximum & minimum*);
– медиана (*median*);
– дисперсия (*variance*);
– асимметрия (*skewness*);
– эксцесс (*kurtosis*).

Кроме того, используются и другие скалярные и векторные характеристики сигналов:

– энергия (*short-time energy*);
– мощность (*power*);
– огибающие сигнала (*envelopes*);
– магнитуда (*short-time average magnitude*);
– частота пересечения нуля (*short-time zero crossing rate*);
– отношение сигнал-шум (*signal-to-noise ratio*);
– автокорреляция (*short-time autocorrelation*);
– спектр (*spectrum*).

1.1 Основные понятия кратковременного анализа сигналов

1.1.1 Идея кратковременного анализа сигналов

Важным свойством речи (и большинства других акустических сигналов) является их нестационарность. В этом случае описание сигналов с помощью средних значений характеристик, использовавшееся в анализе стационарных сигналов, оказывается малоинформативным. Как анализировать и описывать нестационарные сигналы?

Прежде всего, заметим, что сигнал невозможно охарактеризовать по значению в одной точке. Представьте, что вы имеете короткий фрагмент сигнала. Даже на слух невозможно определить, является этот сигнал речью или нет! Нестационарный (речевой) сигнал – это последовательность состояний, каждое из которых имеет протяженность во времени.

Назначение кратковременного анализа сигналов – адекватное описание нестационарных сигналов, позволяющее оценивать их меняющиеся во времени характеристики, описывающие последовательность состояний.

Большинство свойств сигналов проявляется на некотором временном интервале, поэтому и характеристики сигналов также необходимо вычислять на интервалах. Общим принципом КАС является представление нестационарных процессов на отдельных интервалах как локально стационарных, а основным методом КАС является оценка характеристик сигналов на последовательности интервалов.

Данный принцип анализа применим не только для РС, но и для анализа других нестационарных сигналов. Во всех случаях будем использовать термин *кратковременный анализ сигналов*. Кратковременный анализ является инструментом анализа нестационарных временных рядов и сигналов и является базой для большинства алгоритмов обработки РС.

Основные элементы КАС:

– короткие кадры сигнала обрабатываются так, как будто сигнал имеет постоянные характеристики;

– анализ периодически повторяется на протяжении всего сигнала;

– анализируемые кадры анализа могут пересекаться;

– результатом анализа кадра сигнала может быть скаляр или вектор;

– результатом КАС является новая временная последовательность (ряд), который является новым представлением сигнала.

Характеристики являются инструментом анализа, описания и представления сигналов. Стационарный сигнал – это сигнал, характеристики которого не зависят от временной точки, где они измеряются. Стационарные сигналы (процессы) могут быть охарактеризованы с помощью вычисленных по всей реализации средних значений характеристик с использованием *долговременного анализа* (*long-term analysis*).

Реальные сигналы не являются стационарными, но для того, чтобы продолжать использовать характеристики, применяемые при анализе стационарных сигналов, вводят понятие *кратковременной стационарности* (*short-term stationarity*). При этом нестационарные сигналы рассматриваются как локально стационарные. Это означает, что характеристики, оцениваемые на стационарных сигналах, являются достоверными для коротких временных интервалов, а оценки этих характеристик изменяются при переходе от одного интервала анализа к другому. Характеристики сигналов, вычисляемые на интервалах, называются *кратковременными характеристиками* (*short-term characteristics / features*). Важно также, что *свойства сигналов* проявляются на некотором интервале и относятся к этому интервалу. То есть отнесение характеристик к интервалам анализа – это не прихоть, а объективная реальность. Нельзя приписать свойство отдельному отсчету сигнала!

Выбор интервала зависит от особенностей сигнала и задач анализа. Теперь конкретизируем условия оценки параметров для речевых сигналов.

Долговременный анализ характеризует реализацию сигнала (текущую реализацию) в целом, как стационарный временной ряд.

Кратковременный анализ характеризует сигнал на некотором скользящем интервале.

При анализе нестационарных сигналов сосредоточимся на *временном поведении характеристик*.

Поскольку РС – сигнал с относительно медленно меняющимися характеристиками, то можно характеризовать сигнал *средними значениями этих характеристик* на коротких интервалах – кадрах.

Подобно тому, как при дискретизации сигнала он представляется значениями в отдельные моменты времени, характеристики сигнала представляются последовательностью значений в отдельные моменты времени.

1.1.2 Основные предположения КАС для речевых сигналов

При оценке характеристик речевых сигналов на интервалах стационарности возникает проблема выбора длительности этих интервалов.

Некоторые свойства РС меняются относительно медленно (4–10 звуков в секунду), то есть с длительностью 250–100 мс, однако в РС также присутствуют события малой длительности (миллисекунды). Какой интервал анализа использовать? Любой выбор интервала анализа РС приводит к неопределенности оценок параметров:

– малые интервалы анализа (5–20 мс): неопределенность вследствие малого объёма данных, вариаций амплитуды и оценкой основного тона;

– средние интервалы анализа (20–100 мс): неопределенность вследствие быстрых изменение параметров РС на интервале анализа;

– длинные интервалы анализа (100–500 мс): неопределенность вследствие большого числа изменений РС на интервале анализа.

Итак, в кратковременном анализе РС всегда присутствует неопределенность. Выбор интервала анализа РС основан на компромиссном решении.

1.1.3 Базовые принципы кратковременного анализа

Базовыми принципами КАС являются *сегментация, взвешивание и преобразование*. В общем случае оценка вектора параметров в момент k представляется следующей формулой:

$$Q[k] = \sum_{+\infty}^{-\infty} T\{x[n]w[k-n]\},$$

где $Q[k]$ – вектор параметров сигнала $x[n]$ в момент k;

$T\{.\}$ – оператор, определяющий функцию анализа (вычисления параметра);

$w[k-n]$ – весовая функция скользящего окна, задачей которого является выбор сегмента сигнала $x[n]$ и его взвешивание в окрестности точки $n = k$.

Общая схема КАС представлена на рисунке 1.1.

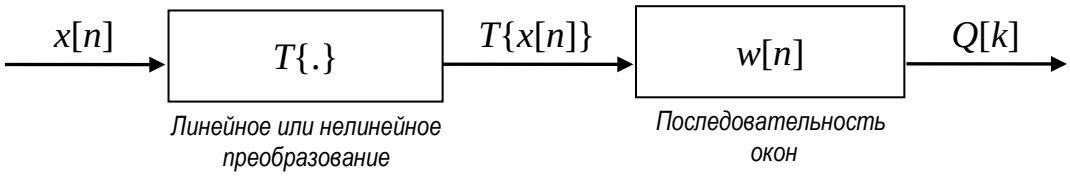
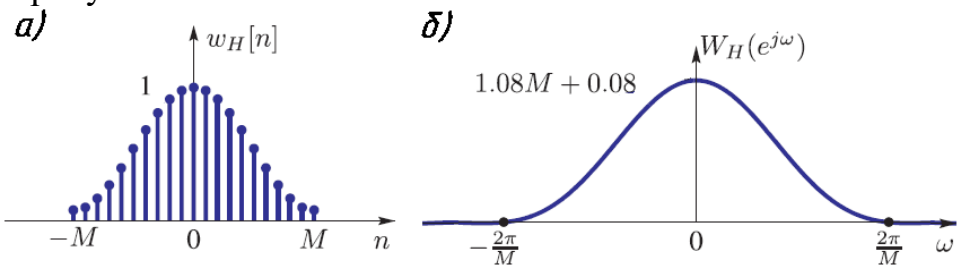


Рисунок 1.1 — Схема кратковременного анализа сигнала

Взвешивание окном $w[n]$ эквивалентно низкочастотной фильтрации, результатом которой является удаление высокочастотных компонент мгновенных оценок параметра и ослабление их случайных вариаций (сглаживание). Форма окна влияет на свойства сглаживающего фильтра. Сравнение окон фильтров низких частот приведено, например в [1].

Пример взвешивающего окна и его амплитудно-частотная характеристика приведены на рисунке 1.2.



а) временно́е представление, б) дискретное преобразование Фурье [2].

Рисунок 1.2 — Взвешивающее окно Хемминга

На рисунке 1.3 приведен пример покадрового выбора и взвешивания сигнала окном Хемминга с 50 %-ным пересечением кадров:

$$\begin{cases} W_H[n] = 0{,}54 + 0{,}46\cos\left(\frac{\pi n}{M}\right), -M \le n \le M \\ \text{иначе} = 0 \end{cases}$$

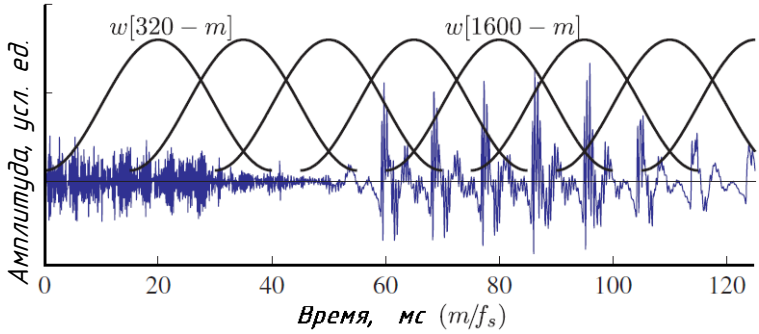


Рисунок 1.3 — Выбор и взвешивание фрагментов речевого сигнала окнами кратковременного анализа [2]

Оператор кратковременного преобразования сигнала в общем случае может давать на выходе значения параметра, векторной характеристики либо значения самого сигнала в интересующем диапазоне частот.

1.1.4 Виды кратковременного анализа сигналов

В зависимости от шага кадров, на которых оцениваются кратковременные характеристики сигнала, возможны следующие варианты КАС:

– «точка за точкой» (*point by point characteristics*);

– «кадр за кадром» с постоянным размером кадров и постоянным шагом анализа;

– «кадр за кадром» с переменным размером кадров и переменным шагом анализа.

1.2 Кратковременные скалярные характеристики

Рассмотрим алгоритмы оценки кратковременных характеристик «точка за точкой» и «кадр за кадром».

1.2.1 Алгоритмы анализа «точка за точкой»

При оценке характеристики «точка за точкой» обычно используют алгоритмы, не требующие большого количества операций. Основные алгоритмы – экспоненциальное сглаживания и сглаживание в скользящем прямоугольном окне.

Экспоненциальное сглаживание

Сглаживание выполняется с применением следующего соотношения:

$$Q[n] = Q[n-1] + \beta \times (T\{x[n]\} - Q[n-1])$$

$$\beta = \frac{1}{\Delta TFs} \approx \frac{2}{Ns},$$

где β – коэффициент сглаживания;

ΔT – временной интервал сглаживания;

Ns – эквивалентная длина интервала сглаживания [3].

Сглаживание скользящим прямоугольным окном

Сглаживание выполняется с применением следующего соотношения:

$$Q[k] = Q[k-1] + \frac{1}{N}\left(T\{x_k[n]\} - T\{x_k[n-N]\}\right) = \frac{1}{N}\sum_{i=0}^{N-1} T\{x_k[kL + n]\},$$

где N – длина кадра;

k – номер кадра;

L – шаг кадра;

$x_k[n]$ – сигнал $x[n]$ на k-м кадре:

$x_0[n] = x[n]$, $x_1[n] = x[L+n]$, $x_2[n] = x[2L+n]$,…. $x_k[n] = x[kL+n]$.

Сглаживание прямоугольным окном реализуют в форме циклического буфера [4].

1.2.2 Кратковременная энергия

Рассмотрим оценку кратковременной энергии с прямоугольным взвешивающим окном [2].

Кратковременная энергия в момент k с прямоугольным окном N:

$$w[n] = \begin{cases} 1 \text{ для } 0 \le n \le N-1 \\ 0 \text{ для } n \ge N \end{cases}$$

вычисляется следующим образом:

$$E(k) = \sum_{n=kL}^{kL-N+1} x^2[n] = x^2[kL-N+1] + \cdots + x^2(kL) = \sum_{n=0}^{N-1} x_k^2[n].$$

Рассмотрим оценку кратковременной энергии в момент k с окном $w[n]$. Представим исходную формулу оценки энергии в виде свертки квадратов амплитуд сигнала с функцией окна. Сигнал после выделения кадра и взвешивания:

$$x_k[n]w[n] = x[n]w[kL-n], \qquad kL - N + 1 \le n \le kL$$

где $w[n]$ – функция взвешивающего окна.

Его кратковременная энергия $E(k)$:

$$E(k) = \sum_{n=-\infty}^{+\infty} (x[n]\, w[kL-n])^2 = \sum_{n=-\infty}^{\infty} x^2[n]\, w^2[kL-n] = \sum_{n=0}^{N-1} x_k^2[n]\, w^2[n].$$

Таким образом, кратковременную энергию можно представить как свертку или линейную фильтрацию, где импульсный отклик линейного фильтра $h[n] = w^2[n]$:

$$E(k) = x_k^2[n] * w^2[n] = x_k^2[n] * h[n].$$

Аналогично вычисляются другие оценки: мощность, энергия, среднее и средневыпрямленное значение. Например, мощность:

$$P(k) = \frac{E(k)}{N} = x_k^2[n] * h_p[n].$$

Формулы основных кратковременных оценок основных параметров сигналов приведены в таблице 1.1.

Таблица 1.1 — Формулы кратковременных оценок параметров сигнала

Параметры сигнала	Вид сглаживания	
	Взвешивающее окно	Экспоненциальное сглаживание
Энергия	$E_x[k] = \sum_{i=0}^{N-1} x_k^2\ [n]w[n]$	E[n] = E[n-1] + β (x²[n]/β – E[n-1])
Мощность	$P_x[k] = \sum_{i=0}^{N-1} x_k^2\ [n]w[n]$	P[n] = P[n-1] + β (x²[n] – P[n-1])
Среднеквадратичное значение	$Rms_k^2 = \sum_{i=0}^{N-1} x_k^2\ [n]w[n]$	Rms²[n] = Rms²[n-1] + β (x²[n] – Rms²[n-1])
Средневыпрямленное значение	$M_x[k] = \sum_{i=0}^{N-1} \lvert x_k\ [n]\rvert w[n]$	M[n] = M[n-1] + β (|x[n]| – M[n-1])
Среднее значение	$\mu_x[k] = \sum_{i=0}^{N-1} x_k\ [n]w[n]$	μ[n] = μ[n-1] + β (x[n] – μ[n-1])

Приведенные формулы могут применяться также при вычислении других характеристик сигнала, например спектра мощности или ковариационной функции.

1.2.3 Сравнение алгоритма экспоненциального сглаживания и сглаживания с прямоугольным окном

Оба алгоритма обрабатывают данные «точка за точкой». Важное отличие алгоритмов заключается в том, что фильтр экспоненциального сглаживания (ФЭС) имеет бóльшую ширину полосы пропускания и бóльшее время реакции:

$$(\Delta F \times Ta)_{\text{ФЭС}} \gg (\Delta F \times Ta)_{\text{Прямоугольное окно}}$$

В результате эффективность сглаживания у ФЭС существенно хуже, чем у прямоугольного окна. Свойства окон иллюстрируются рисунками 1.4 и 1.5.

На рисунке 1.4 показаны скользящие оценки среднеквадратичного значения (СКЗ) фрагмента речевого сигнала, вычисленные с прямоугольным и экспоненциальным окном. Из рисунка следует, что прямоугольное окно обеспечивает более крутые временные фронты и меньшие вариации СКЗ на тональных участках РС.

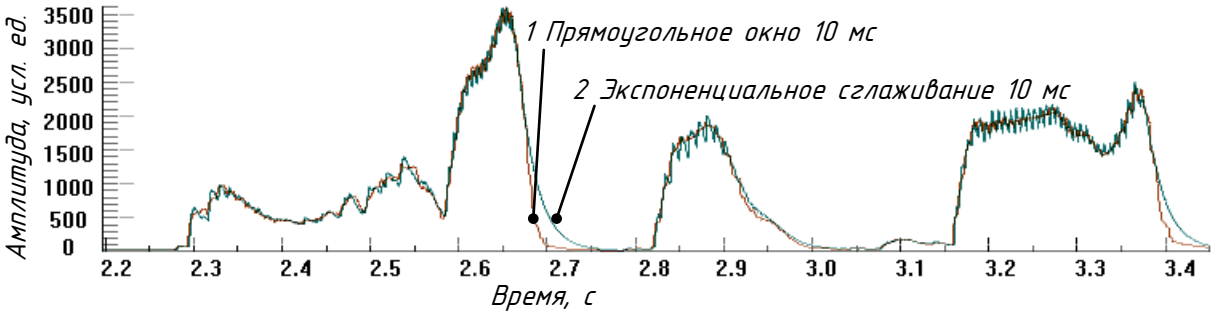


Рисунок 1.4 — Скользящие оценки среднеквадратичного значения

На рисунке 1.5 показаны спектры широкополосного сигнала до и после низкочастотной фильтрации (оценки среднего значения) прямоугольным окном и ФЭС. Из рисунка следует, что прямоугольное окно обеспечивает существенно лучшее подавление сигнала вне границы полосы пропускания. Отметим также, что применение специальных (не прямоугольных) окон приводит к значительному дополнительному подавлению компонент временной последовательности оценок параметров за границей полосы пропускания.

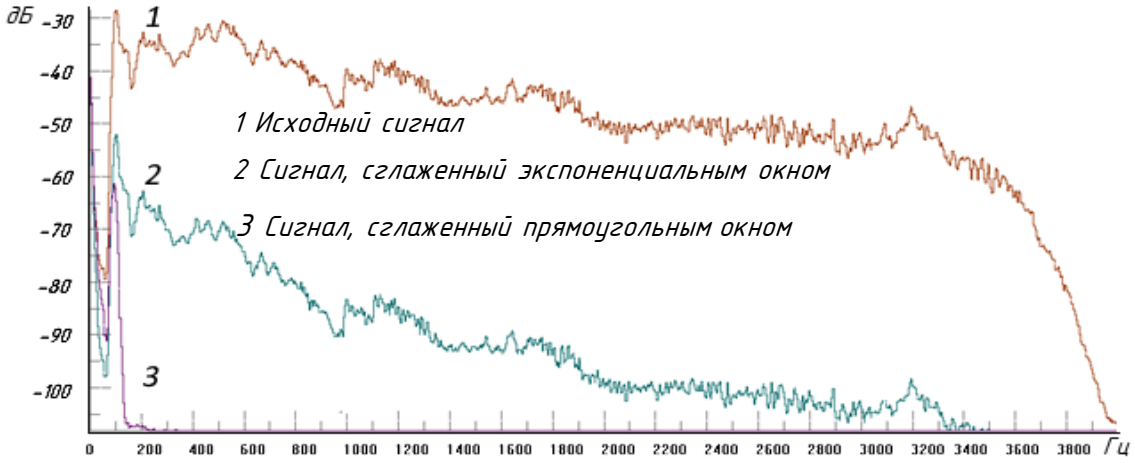


Рисунок 1.5 — Средние спектры сигналов

1.2.4 Задание параметров КАС

Параметрами КАС являются размер окна и шаг (сдвиг) окна от кадра к кадру. Выбор их значений зависит от решаемой задачи.

Выбор размера окна

Рассмотрим вопрос выбора размера окна применительно к оценке мощности РС. Для оценки текущей *мощности тонального* РС размер окна должен охватывать несколько периодов импульсов тонального РС (основного тона).

Минимальные значения частоты основного тона (ОТ) составляют приблизительно 80 Гц. В этом случае период ОТ равен 12,5 мс. Размер окна анализа должен быть не менее двух периодов, т.е. 25 мс. Для большинства тональных звуков достаточным является размер окна 20 мс.

Рассмотрим пример оценки СКЗ тонального речевого сигнала с окнами разного размера. На рисунке 1.6 приведены графики СКЗ тонального РС, вычисленные с прямоугольными окнами длительностью 2 и 30 мс.

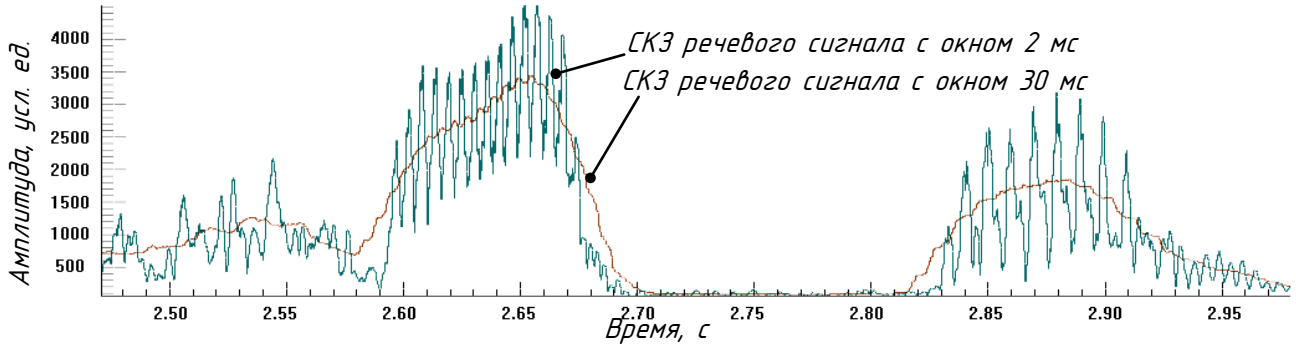


Рисунок 1.6 — СКЗ речевого сигнала с окнами 2 и 30 мс

Из графиков видно, что окно длительностью 30 мс позволяет получить огибающую мощности тонального сигнала, в то время как с окно 2 мс огибающая характеризует мощность отдельных импульсов РС, то есть не позволяет решить поставленную задачу.

Рассмотрим другой пример – оценку огибающей мощности импульса. На рисунке 1.7 приведена осциллограмма импульса и графики оценки СКЗ, рассчитанные с длительностью окна 2, 4 и 8 мс.

Импульсный сигнал включает в себя широкополосный набор спектральных компонент. Различные графики описывают огибающие мощности (или СКЗ) одного и того же сигнала в разных частотно-временных областях: «размер окна»×«ширина спектра» $\Delta T \times \Delta F$. При этом длительность отклика составляет 2, 4, 8 мс, полоса спектра мощности импульса при оценке СКЗ составляет 500, 250 и 125 Гц.

Выбор размера окна анализа зависит от решаемой задачи. Если нас интересует точное положение импульса, то следует использовать окно малой длительности, если интересует область импульса с окружением, то окно должно быть больше.

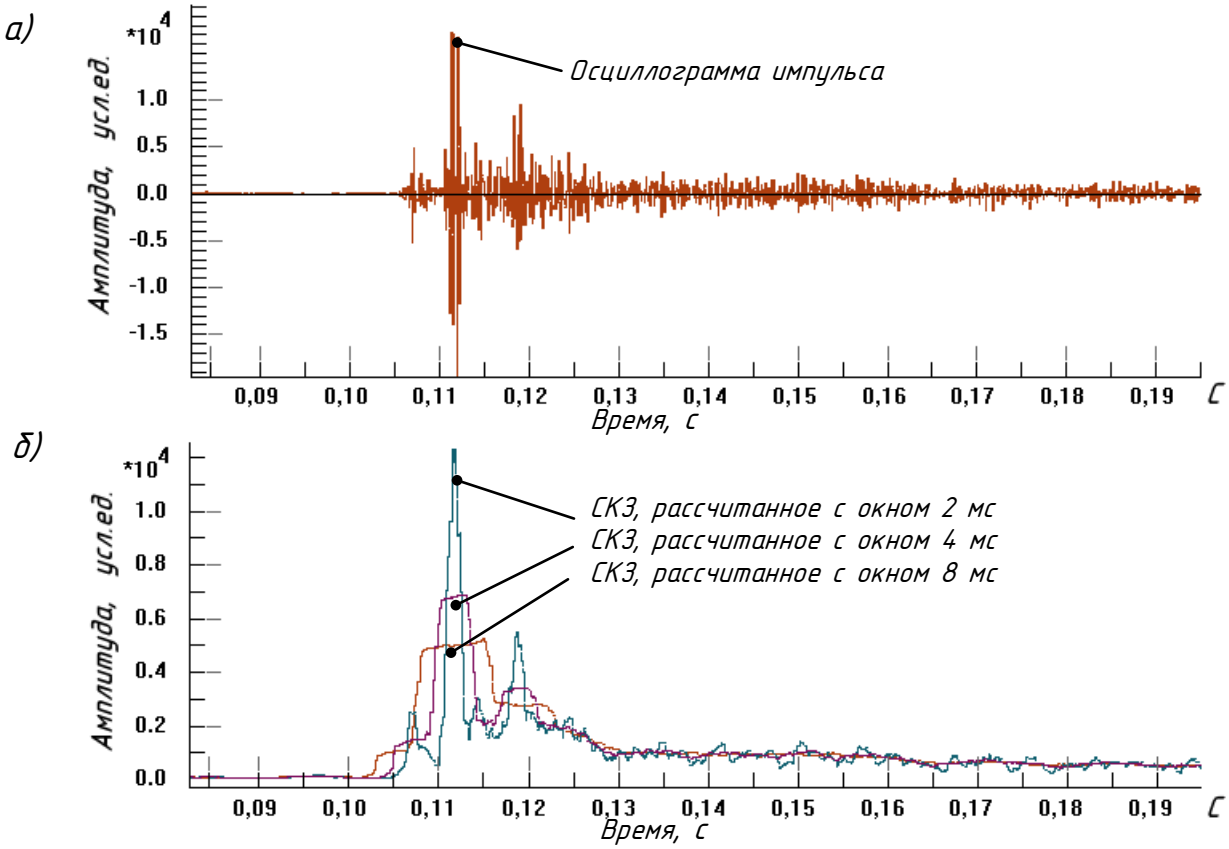


Рисунок 1.7 — Зависимость СКЗ импульса от размера окна

Выбор шага анализа

Динамика значений параметра от кадра к кадру характеризуется спектром значений оценок параметра на кадрах.

С одной стороны, временной шаг анализа и связанная с ним частота выборки, должны соответствовать ширине этого спектра. С другой стороны, размер кадра N (то есть его длина ΔT) ограничивает ширину спектра ΔF:

$$\Delta F \approx \frac{1}{\Delta T} = \frac{1}{NT} = \frac{F_s}{N},$$

где T – шаг дискретизации сигнала;

Fs – частота дискретизации сигнала;

ΔT – длина кадра.

Шаг сдвига кадров L определяет частоту кадров Fk:

$$Fk = \frac{F_s}{L}.$$

Шаг выборки значений оценки параметра должен быть таким, чтобы избежать эффекта наложения спектров (*aliasing*), то есть частота следования кадров должна превосходить удвоенную ширину спектра оценок параметра: $Fk \geq 2\Delta F$.

То есть $\frac{F_s}{L} \geq 2\frac{F_s}{N}$, отсюда $L \leq \frac{N}{2}$.

В качестве иллюстрации взаимосвязи длины кадра анализа и ширины спектра параметра рассмотрим следующий пример оценки СКЗ сигнала. Амплитуда огибающей широкополосного шума промодулирована на разных временных интервалах с частотами 0,625 – 1,2… – 2,5…5…10…12,5 Гц.

На рисунке 1.8 приведены графики оценки СКЗ сигнала с окнами длительностью 5 и 80 мс.

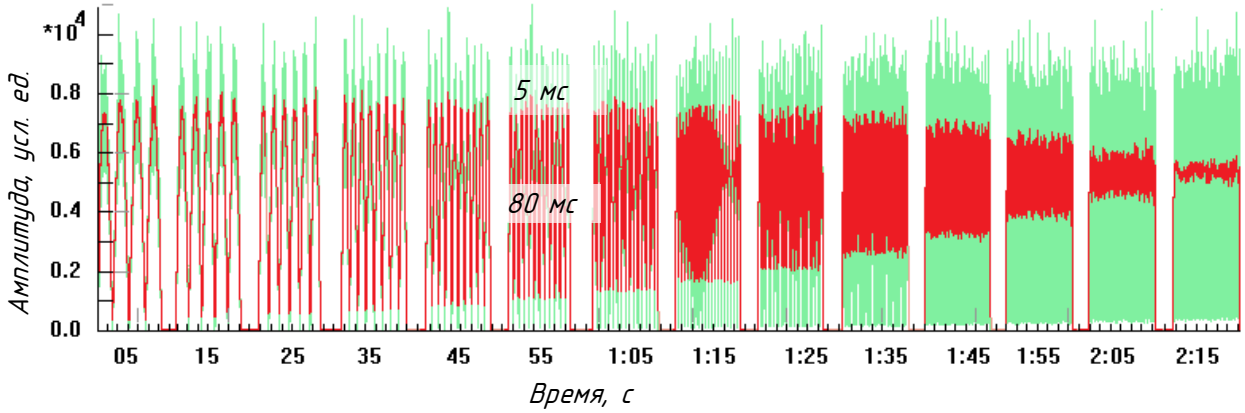


Рисунок 1.8 — Оценки СКЗ сигнала, рассчитанные с прямоугольными окнами 5 и 80 мс

Из рисунка следует, что по мере увеличения размера окна оценка вариаций СКЗ на частоте 12,5 Гц ослабляется, что свидетельствует о подавлении её высокочастотных компонент. Более детально данный эффект иллюстрируется рисунком 1.9.

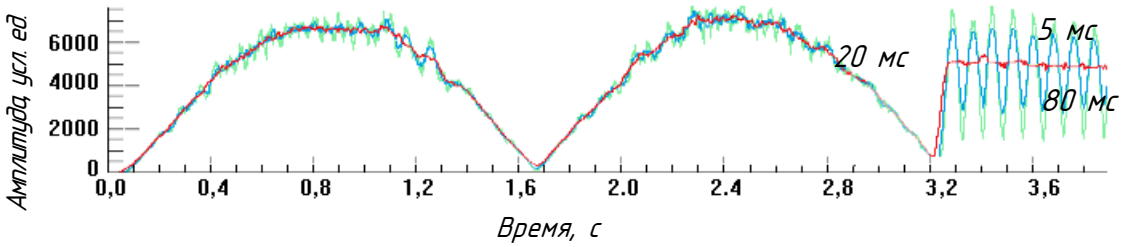


Рисунок 1.9 — Оценки СКЗ сигнала, рассчитанные с окнами анализа 5, 20 и 80 мс

Из рисунка следует, что в случае окна ΔT=20 мс получим $\Delta F \approx 50$ Гц $\gg$12,5 Гц. В этом случае динамика оценки СКЗ соответствует реальной модуляции сигнала. Для ΔT = 80 мс, $\Delta F \approx$ 12,5 Гц модуляция сигнала не отображается.

При усреднении с использованием алгоритма экспоненциального сглаживания эквивалентная длина окна сглаживания рассчитывается по формуле: $Ns \approx 2/\beta$, где β – коэффициент сглаживания.

1.2.5 Влияние частоты дискретизации на кратковременную энергию речевого сигнала

Частота дискретизации ограничивает верхнюю границу частотного диапазона сигнала. Для широкополосного сигнала (например, белого шума) уменьшение частоты дискретизации приведет к пропорциональному уменьшению его мощности, но для РС будет иначе. Основная часть энергии РС сосредоточена в диапазоне низких частот (НЧ). Уменьшение частоты дискретизации приводит к исчезновению высокочастотных компонент, дающих незначительный вклад в общую энергию сигнала.

Различие может проявиться лишь на широкополосных звуках с высоким уровнем энергий в диапазоне высоких частот (ВЧ).

1.2.6 Частота пересечения нуля

Частота пересечений нуля (ЧПН, *zero crossing rate, ZCR*) определяется как среднее число пересечений нуля на отсчет согласно следующей формуле [5]:

$$Z_k = \frac{1}{2N} \sum_{n=-\infty}^{+\infty} |\operatorname{sgn}(x[n]) - \operatorname{sgn}(x[n-1])|\ w[k-n]\,,$$

где N – длина окна анализа;

$w[n]$ – прямоугольное окно: $w[n]=1$ для $0 \le n \le N-1$, $w[n] = 0$ за пределами окна,

$$\operatorname{sgn}(x) = \begin{cases} 1, x > 0 \\ 0, \ x = 0 \\ -1, \ x < 0 \end{cases}.$$

На рисунке 1.10 приведена схема кратковременной оценки ЧПН.

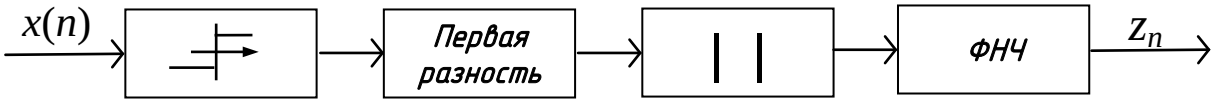


Рисунок 1.10 — Схема кратковременной оценки ЧПН [5]

На практике используется нормализованная оценка ЧПН – число пересечений нуля на интервале 1 с, определяемая следующим соотношением:

$$Zo_k = Fs\, Z_k.$$

Нормализованная оценка приблизительно инвариантна к частоте дискретизации сигнала, поскольку число N отсчетов сигнала на интервале 1 с меняется пропорционально частоте дискретизации Fs.

Свойства среднего числа пересечений нуля:

– среднее число пересечений нуля — простая мера частотного состава сигнала;

– синусоида с частотой Fo имеет $2Fo$ пересечений в секунду;

– нормализованная мера ЧПН не зависит от частоты дискретизации сигнала;

– для корректной оценки ЧПН необходимо убрать постоянную составляющую и наводки, то есть осуществить полосовую фильтрацию сигнала;

– оценка ЧПН в отдельных частотных полосах РС позволяет грубо оценить доминирующую частоту спектра РС в соответствующей полосе.

Пример

Частота пересечения нуля сигнала $\sin(2\pi Fo\ n/Fs)$, Fo = 432 Гц для сигналов с частотами дискретизации 16 и 8 кГц будет одинаковой: $Zo \approx 864$ Гц.

Пример

В то время как ЧПН тонального сигнала не зависит от частоты дискретизации, в случае РС это может быть не так.

На рисунке 1.11 приведена спектрограмма РС (Fs = 16 кГц), его среднее число пересечений нуля и для сигнала с уменьшенной вдвое частотой дискретизации Fs = 8 кГц.

Из рисунка следует, что на тональных участках сигнала ЧПН не меняется, а на шумовых широкополосных уменьшается вдвое с уменьшением вдвое величины *Fs*.

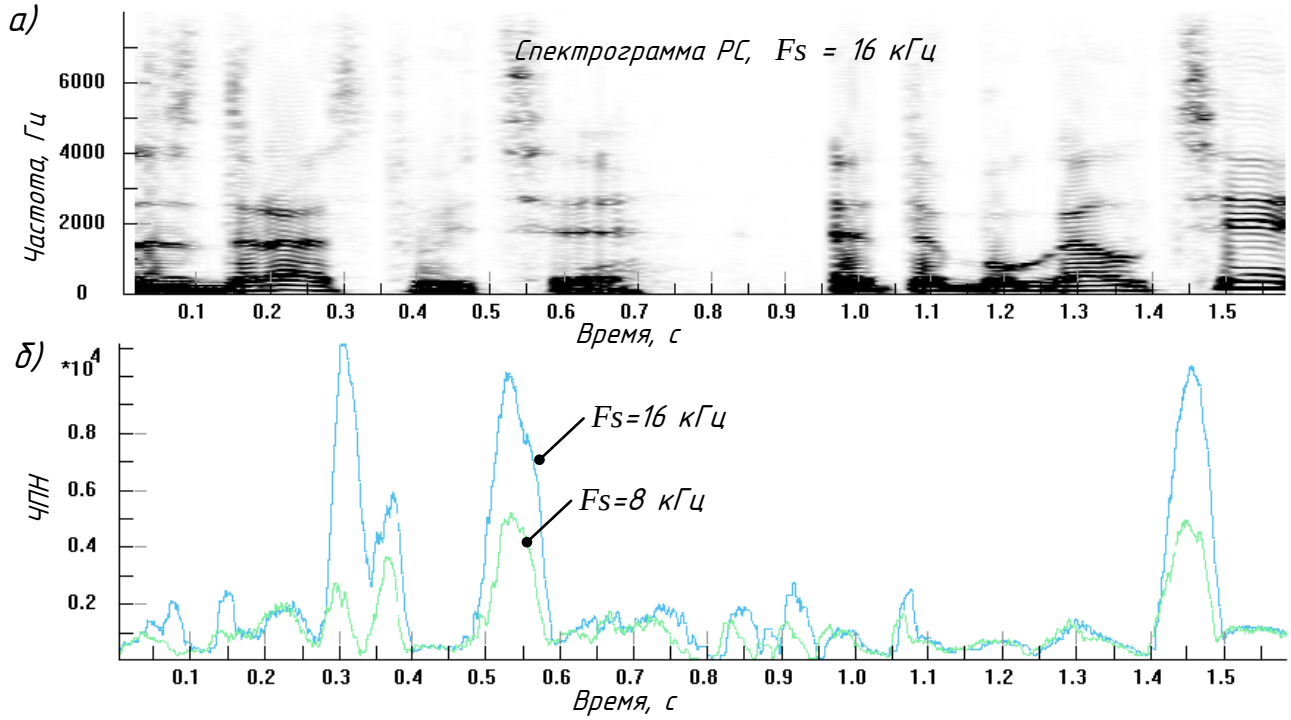


Рисунок 1.11 — а) спектрограмма речевого сигнала (*Fs*=16 кГц);
б) частота пересечения нуля сигнала с разной частотой
дискретизации сигнала (*Fs* = 16 и 8 кГц)

Пример
Исследовались оценки ЧПН речевого сигнала *Fs* = 8 кГц, отношением сигнал-шум (ОСШ) 30 дБ с различными размерами окон анализа.
Осциллограмма и график мощности сигнала с окном 20 мс приведены на рис. 1.12 а).
Оценка частоты пересечения нуля с окнами 20 и 80 мс приведена на рис. 1.12 б).

Из приведенных рисунков следует, что максимумы мощности сигнала относятся к участкам тональных звуков, а максимумы и ЧПН — к участкам широкополосных шумоподобных звуков.

На тональных участках РС происходит понижение ЧПН, а на участках фонового шума происходит некоторое увеличение ЧПН.

Значительное влияние на оценку ЧПН может оказывать смещение сигнала, т.е. добавление к сигналу некоторой постоянной величины (B), превосходящей по величине амплитуду шумового фона.

На рисунке 1.13 приведены графики ЧПН (окно 20 мс) центрированного речевого сигнала $x[i]$ и сигнала со смещением B = 200. $y[i] = x[i] + B$.

Из рисунков видно, что участки с небольшой мощностью (фон или широкополосные компоненты РС с амплитудой меньше 200) перестают вносить вклад в ЧПН.

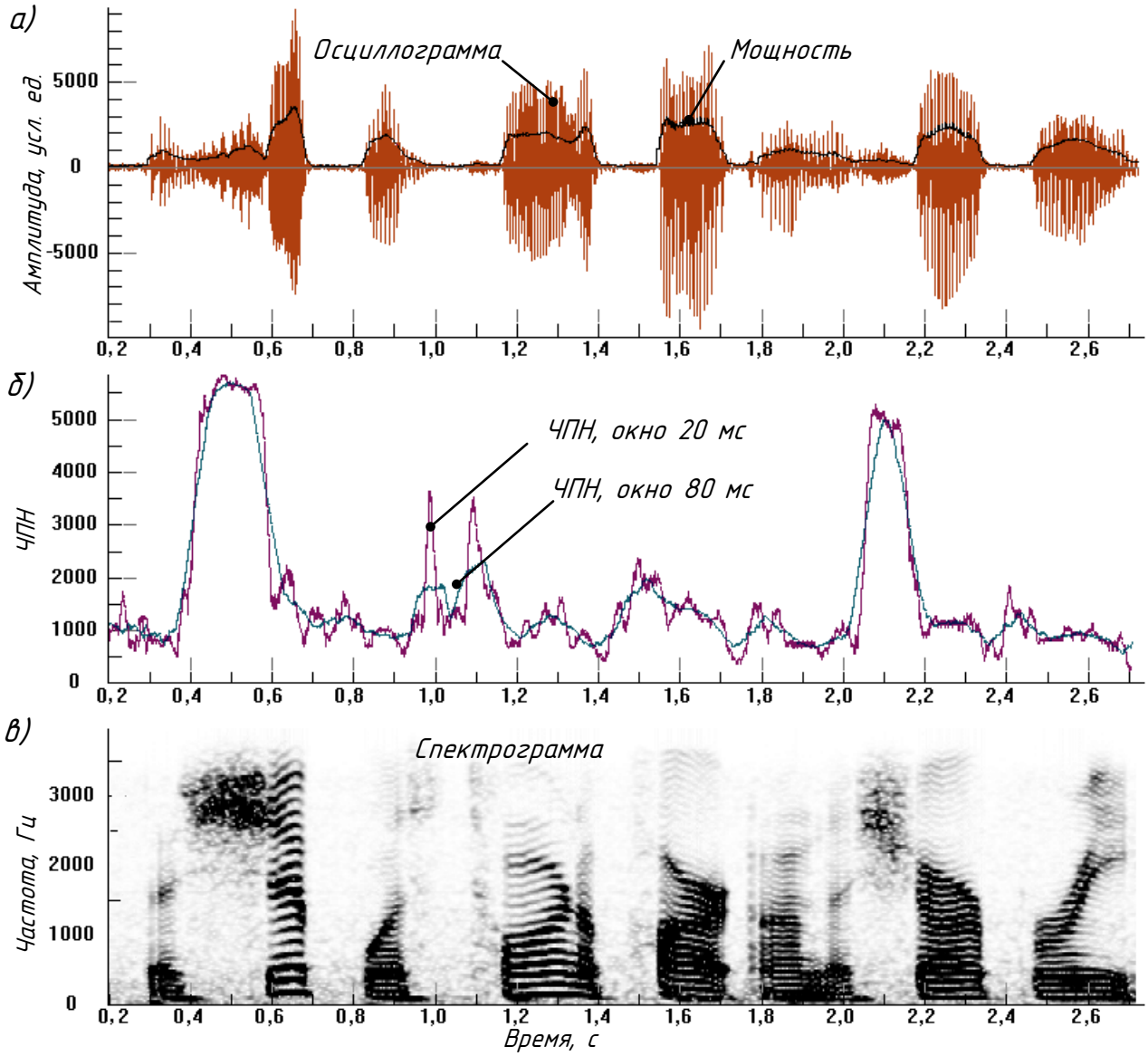


Рисунок 1.12 — а) осциллограмма и мощность РС с окном 20 мс;
б) частота пересечения нуля с окнами 20 и 80 мс; в) спектрограмма

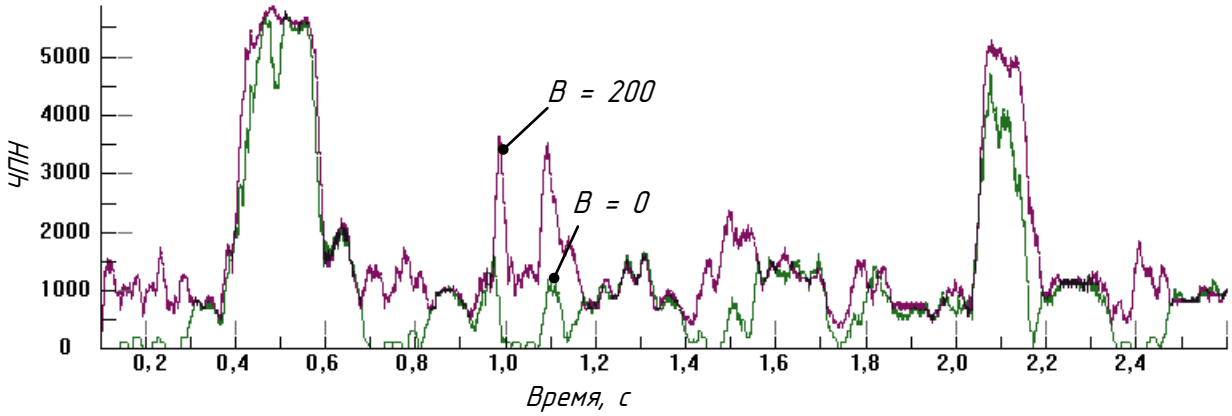


Рисунок 1.13 — Частота пересечения нуля с окном 20 мс речевого сигнала с нулевым средним и со смещением 200

1.3 Кратковременный корреляционный анализ

Автоковариационная функция (АКФ) используется для детектирования периодичности сигналов и является основой многих методов анализа РС.

Кратковременная автоковариация (*short time autocorrelation function, STACF*) определяется как функция последовательности $x_k[n] = x[n]w[k - n]$, формируемой окнами, сдвинутыми по времени:

$$C_k[m] = \sum_{n=-\infty}^{+\infty} x_k[n]\, x_k[n+m] =$$

$$= \sum_{n=-\infty}^{+\infty} x[n]w[k - n]x[n + m]w[k - n - m] =$$

$$= \sum_{n=-\infty}^{+\infty} x[n]\, x[k - m]\, \acute{w}_m[n][k - n],$$

где $\acute{w}_m[n] = w[n]w[n + k]$.

$C_k[m]$ является функцией 2-х аргументов – положения окна k и задержки m. Положение окна определено в номерах отсчетов сигнала. В представлении индекса номеров окон, следующих с шагом L, можно записать:

$$x_k[n] = x[n]w[kL - n].$$

АКФ содержит в себе информацию об энергии сигнала, поскольку:

$$E_k = \sum_{n=-\infty}^{+\infty} (x[n]\, w[k - m])^2 = C_k(0).$$

Коррелограмма – графическое представление последовательности АКФ от кадра к кадру. Темным обозначаются положительные значения, светлым – отрицательные функции автокорреляции. Пример коррелограммы РС показан на рисунке 1.14.

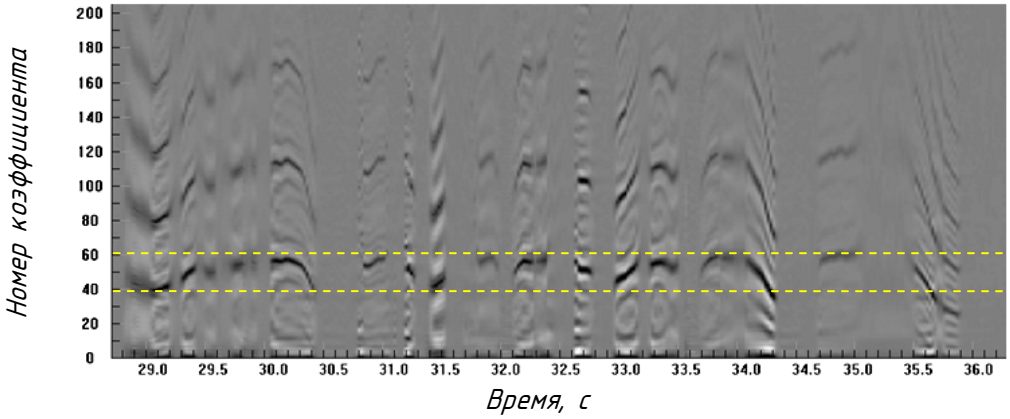


Рисунок 1.14 — Пример коррелограммы речевого сигнала

Обычно коррелограмма представляется в пространстве номеров задержек (соответствующих задержкам отсчётов по времени). Тогда на коррелограмме можно наблюдать корреляционные связи между отсчетами сигнала с разными задержками, например соответствующие периодической структуре сигнала.

Для оценки временных задержек необходимо знать частоту дискретизации сигнала. Так, на рисунке 1.14 наблюдаются максимумы в интервале 40–60 отсчетов, соответствующие для сигнала с частотой дискретизации 8 кГц временным задержкам 5–7,5 мс, что соответствует частотам основного тона 200–130 Гц.

1.4 Применение кратковременного анализа сигналов

КАС широко применяется в задачах детектировании событий, анализа, классификации, сегментации и обработки сигналов. Рассмотрим два примера.

1.4.1 Детектирование импульсных помех

В обработке сигналов для детектирования границ областей с разной мощностью можно применить модуль градиента или его квадрат. Например, детектировать границы импульса можно на основе следующего соотношения:

$$\frac{< (x[t] - x[t-1])^2 >}{T^2 < x[t]^2 >} > THR,$$

где THR –порог детектирования;

$< >$ – оператор усреднения по времени;

T – шаг дискретизации.

Усреднение по времени выполняется на основе алгоритма экспоненциального сглаживания с эффективной длиной окна 5 мс.

1.4.2 Автоматическая регулировка усиления

Для выравнивания уровня сигнала на разных участках используется автоматическая регулировка усиления (АРУ). Коэффициент усиления может быть вычислен следующим образом:

$$G[t] = (Mo/Mx[t]),$$

где $Mx[t]$ – средневыпрямленное значение сигнала;

Mo – заданная (целевая) величина средневыпрямленного значения.

Средний уровень сигнала приводится к заданному уровню с использованием следующего преобразования: $y[t] = G[t]\,x[t]$.

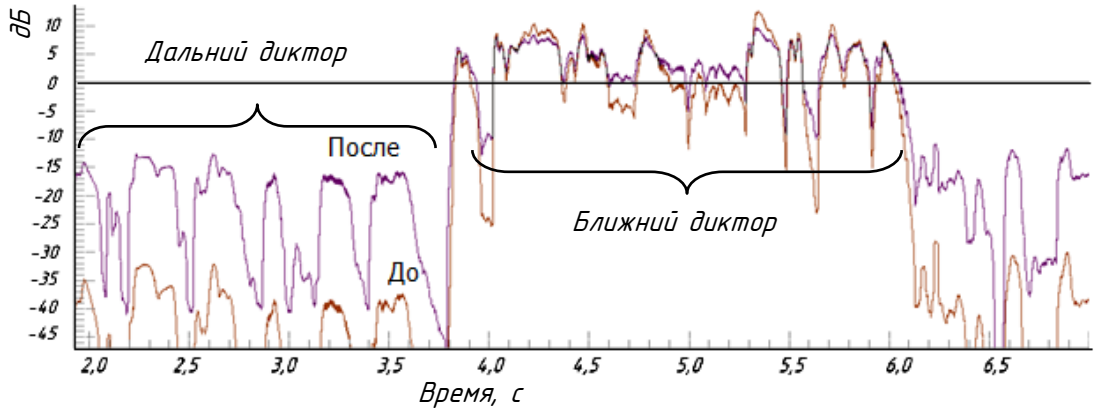


Рисунок 1.15 — Мощность сигнала до и после обработки

На рисунке 1.15 показаны графики мощности (шкала дБ) речевых сигналов дальнего диктора (2,0–3,7 с) и ближнего диктора (3,7–6,0 с) в телефонной линии до и после выравнивания их уровней с помощью АРУ.

Свойства автоматической регулировки усиления зависят от скорости адаптации на участках подъёма и участках уменьшения мощности сигнала (*attack rate, release rate*).

В результате регулировки усиления мощность речевого сигнала дальнего диктора увеличилась на 20 дБ.

Литература

1. Рабинер Л., Гоулд Б. Теория и применение цифровой обработки сигналов // М.: Мир, 1978. — 848 с. (Гл. 12. Цифровая обработка речевых сигналов).

2. Rabiner L. R. and Schafer R. W. Introduction to Digital Speech Processing // Foundations and Trends in Signal Processing. Vol. 1, No.1–2 (2007) 194 P.]

3. Столбов М.Б., Основы анализа и обработки речевых сигналов – СПб.: НИУ ИТМО, 2021. – 101 с.

4. Отнес Р., Эноксон Л. Прикладной анализ временных рядов. Основные методы // М.: Мир, 1982. — 432 с.

5. Рабинер Л. Р., Шафер Р.В. Цифровая обработка речевых сигналов // М.: Радио и связь, 1981. – 496 с. (Гл. 4. Методы обработки речевых сигналов во временной области).

6. Lerch A. An introduction to Audio Content Analysis // John Wiley & Sons, Inc., 2012.

Вопросы и задачи

1. Частота дискретизации сигнала равна 16 кГц. Кадры размером 512 отсчетов берутся с пересечением 75 %. Какова частота кадров?

2. Чем определяются размер, шаг и частота кадров анализа РС?

3. Каковы основные кратковременные характеристики сигнала?

4. Входной сигнал $x[n]$ сегментируется в последовательность кадров. Выразите отсчеты сигнала $xk[n]$ на непересекающихся (идущих встык) кадрах размера N через значения $x[n]$ (k – индекс кадра).

5. Запишите кратковременную скользящую оценку коэффициента корреляции двух сигналов в форме ФЭС.

6. Частота синусоиды равна $F0$ = 1000 Гц, частота дискретизации сигнала Fs = 10 кГц. Чему равна частота пересечения нуля?

Глава 2
Кратковременный спектральный анализ

Средние значения параметров характеризуют постоянство, локальные – изменчивость во времени. *Спектральный анализ* (СА) основан на предположении о повторяемости (периодичности) процессов.

Спектральный анализ – это моделирование процесса (сигнала) в форме суммы периодических компонент, то есть в виде повторяющихся событий. Каждый период повторения характеризуется собственной частотой.

Важной особенностью СА является возможность разложения сигнала на линейно независимые (ортогональные) компоненты, их преобразование и последующее восстановление (синтез сигнала) с помощью суммирования этих компонент.

2.1 Кратковременный спектрально-временной анализ сигналов

2.1.1 Понятие кратковременного спектрального анализа

Спектральный анализ – это инструмент анализа и представления процессов, позволяющий обнаруживать периодические (повторяющиеся) события и оценивать их значимость на разных частотах. Использование спектра сигнала (разложения по базису ортогональных периодических функций) учитывает «волновую природу» сигнала.

Кратковременный спектральный анализ (КСА) – это теория, методы и инструменты анализа и обработки нестационарных сигналов в спектрально-временной области (*time-frequency domain*).

КСА широко применяется при решении многих задач, связанных с акустическими сигналами, в частности:

– автоматическое распознавание речи;
– распознавание дикторов;
– анализ акустических волн в геофизике и гидрофизике;
– акустический контроль;
– детектирование акустических событий;
– анализ музыкальных сигналов;
– анализ медицинских и биофизических сигналов.

Исчерпывающее описание различных методов и алгоритмов КСА изложено в монографии [1] и обзоре [2].

Представление сигналов в частотно-временном пространстве может выполняться с использованием анализа Фурье (*FT view*) или с помощью фильтрации сигнала гребенкой фильтров (*filter view*).

Анализ Фурье приводит к равномерной по частоте шкале спектра сигнала, которая может преобразовываться в неравномерную, например, мел-шкалу и другим. КСА на основе гребенки фильтров позволяет представить сигнал в выбранной шкале частот.

Основная идея КСА на основе анализа Фурье (*short-time fourier transform STFT*) состоит в следующем. Последовательность отсчетов на кадрах сигнала $\mathbf{X}[k] = \mathbf{X}_k$ взвешивается оконной функцией и преобразуется с помощью дискретного преобразования Фурье (ДПФ) в последовательность кратковременных *комплексных спектров*:

$$\{\mathbf{X}[k]\} \rightarrow DFT \rightarrow \{X(n,k)\},$$

где k – индекс кадра;

n – индекс частоты n-й полосы спектра f_n.

Последовательность отсчетов спектра $\{X(n,k)\}$ можно рассматривать как комплексный сигнал в частотной полосе f_n, взятый с частотой следования кадров. При этом временной сигнал может быть реконструирован из последовательности кратковременных комплексных спектров с помощью процедуры *перекрытия и суммирования* [3], другое название – процедура *перекрытия и сложения* [4].

На основе последовательности комплексных спектров можно вычислить другие кратковременные спектральные характеристики:

– *периодограмма*: $Ix(n,k) = |X(n,k)|^2$;

– *амплитудный спектр*: $Ax(n,k) = |X(n,k)|$;

– *фазовый спектр*: $\Phi x(n,k) = \arctan[\mathrm{Im}\{X(n,k)\} / \mathrm{Re}\{X(n,k)\}]$;

– *кратковременный спектр мощности*: $Pxx(n,k)$.

Аналогичным образом для двух сигналов определяются *кратковременные кросс-спектральные характеристики*.

Мгновенный кросс-спектр (кросс-периодограмма):

$$Ixy(n,k) = X^*(n,k)Y(n,k),$$

Кратковременный кросс-спектр:

$$Pxy(n,k) = < X^*(n,k)Y(n,k) >_T,$$

где $<..>_T$ – обозначение усреднения по времени.

Кратковременная комплексная функция когерентности (*coherence function*):

$$\Gamma xy(n,k) = \frac{Pxy(n,k)}{[Pxx(n,k)Pyy(n,k)]^{1/2}]}.$$

Кратковременная амплитудная функция когерентности (*magnitude-squared coherence function, MSC*):

$$Gxy(n,k) = \Gamma xy^2(n,k) = \frac{|Pxy(n,k)|^2}{[Pxx(n,k)Pyy(n,k)]}.$$

Соотношение между кратковременным спектром и кратковременной функцией ковариации аналогично соотношению между долговременными характеристиками:

$$Cxx(m,k) = IFT\{Pxx(n,k)\} \leftrightarrow Pxx(n,k) = DFT\{Cxx(m,k)\},$$
$$Cxy(m,k) = IFT\{Pxy(n,k)\} \leftrightarrow Pxx(n,k) = DFT\{Cxx(m,k)\},$$

где m – индекс временной задержки.

2.1.2 Спектрограмма

Спектрограмма – удобный и самый распространенный инструмент для анализа и визуализации речевых и многих других акустических сигналов.

Спектрограмма – двумерный спектрально-временной образ сигнала, получаемый с помощью кратковременного преобразования Фурье или фильтрации.

На рисунках 2.1 и 2.2 приведены различные способы представления спектрограмм.

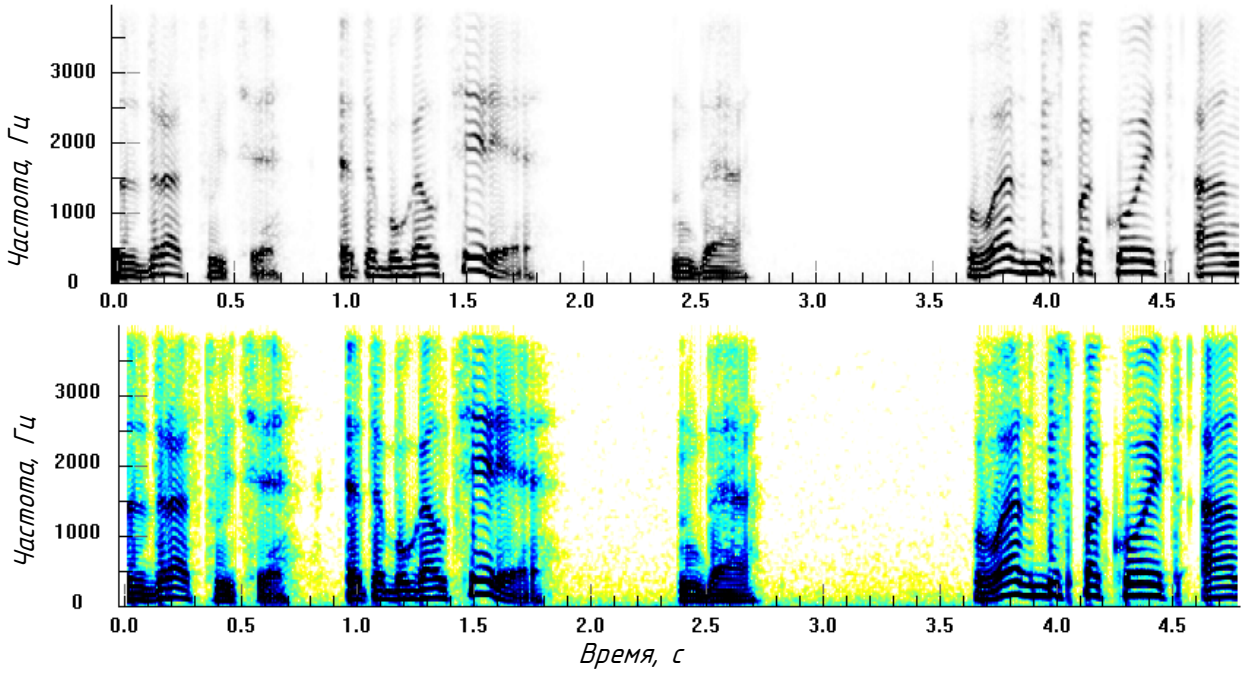


Рисунок 2.1 — Представление спектрограммы оттенками серого и цвета

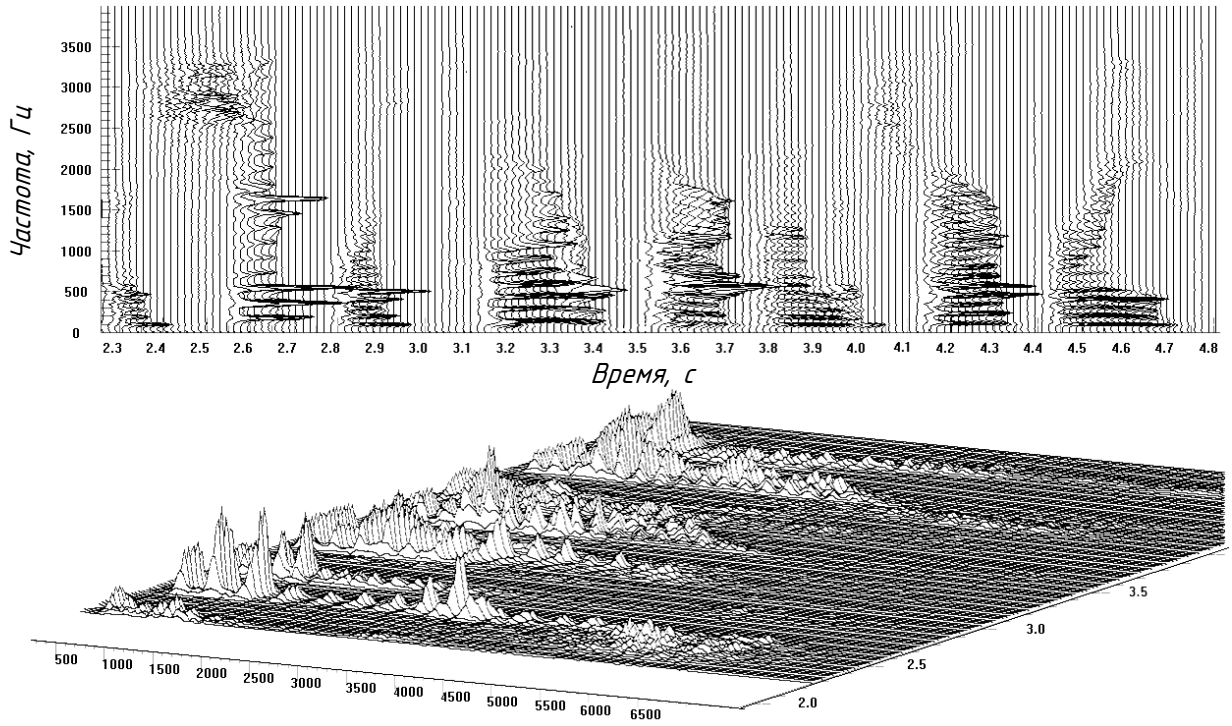


Рисунок 2.2 — Представление спектрограммы отклонением вправо
и аксонометрией

Спектрограмма является формой графического представления результатов КСА в координатах «время-частота-амплитуда». Отображение амплитуд спектров соответствует слуховому восприятию, поскольку человеческий слух не чувствителен к фазе сигнала.

В математическом смысле спектрограмма является матрицей значений $S[n,k]$, где n – дискретный индекс частоты, k – индекс времени (может совпадать с индексом отсчетов сигнала в случае полосовой фильтрации или соответствовать индексу кадров).

Амплитуды спектрограммы могут кодироваться оттенками серого, цветом, графиком огибающей и другими способами.

2.1.3 Параметры кратковременного спектрального анализа

Основными параметрами КСА являются *частота дискретизации сигнала, размер кадра (окна) анализа, шаг окна* (*частота кадров*). Задание значений этих параметров зависят от решаемой задачи и типа сигнала.

Рассмотрим задание этих параметров для речевых сигналов, исходя из временных и частотных свойств РС.

Частота дискретизации определяется свойствами РС, свойствами слуха и решаемой задачей. Частотный диапазон некоторых звуков речи достигает 7 октав и более и выходит за границы диапазона слуха (20 кГц). Д для их представления необходима высокая частота дискретизации (более 40 кГц). Однако в практической плоскости вопрос формулируется иначе: какая частота дискретизации является достаточной для решения конкретной задачи, в частности для прослушивания речи.

Разные частотные диапазоны имеют разное значение для разборчивости речи. Для качественного представления речевого сигнала достаточно частоты дискретизации 44 кГц, для речевого сигнала, передаваемого по телефонному каналу, достаточным является диапазон (300–3500 Гц), то есть частота дискретизации 8 кГц. Для некоторых приложений можно использовать еще меньшую частоту.

Размер кадров анализа. Рассмотрим выбор размера кадра с позиций частотного и временного представления РС.

Во-первых, результатом СА должно быть выделение тональных компонент основного тона (ОТ) речи: от 80 Гц и выше. Для представления спектров таких сигналов частотное разрешение ΔF должно составлять 25..40 Гц. Между дискретным разрешением по частоте и размером кадра анализа имеется отношение неопределенности:

$$\Delta F \le \frac{1}{Nf \times T} = \frac{F_s}{Nf},$$

где Nf – размер кадра анализа.

Отсюда получаем необходимый размер кадра анализа: $Nf \ge Fs/\Delta F$.

Во-вторых, РС содержит события длительностью ΔT = 5–40 мс. Для различения таких событий понадобится кадр размером Nt, такой, что

$$\Delta T \le Nt \times T = \frac{Nt}{Fs}.$$

Отсюда получаем $Nt \le \Delta T \times Fs$.

Очевидно, имеется противоречие между разрешением по времени и частоте, которое не может быть устранено в рамках классического спектрального анализа вследствие принципа неопределенности:

$$\Delta T \times \Delta F \ge 1\,.$$

Для описания различных звуков РС необходимо использовать разную длительность интервала (окна) наблюдения: от 5 до 40 мс.

Однако обычно для анализа сигнала используются кадры фиксированного размера. Размер кадра анализа задается на основе компромисса между требованиями описания разных свойств сигнала.

Обычно длительность кадров анализа РС составляет 20–30 мс. При этом в случае применения быстрого преобразования Фурье (БПФ) размеры окна берут кратными степени двойки.

Типичные размеры кадров КСА для различных частот дискретизации сигнала *Fs* приведены в таблице 2.1.

Таблица 2.1 — Размеры кадров КСА для различных частот дискретизации

Частота *Fs*, кГц	Размер кадров, отсчеты		
	Nf	*Nt*	*Nfft*
8–10	250..400	50..400	256
20	500..800	100..800	512
40	1000..1600	200..1600	1024
более 40	более 1600	более 1600	2048

Частота следования кадров анализа. Третьим параметром КАС является шаг следования кадров и связанная с ним частота кадров анализа (*frame rate*):

$$Fk = \frac{F_s}{L},$$

где L – шаг кадров анализа;

Fk – частота кадров.

Динамика речевых сигналов характеризуется последовательностью значений параметров сигнала на кадрах. Применительно к спектральному анализу такими параметрами являются амплитуды спектров, последовательность которых описывается *временными огибающими спектров* в разных частотных полосах.

Размер кадра ограничивает динамику огибающих, поэтому при выбранном размере кадра нет смысла брать окна анализа со сдвигом в один отсчет.

Достаточным является шаг окна, равный 25–50 % от величины окна анализа. Обычно для частотного анализа РС шаг кадров берут равным половине кадра анализа:

$$L = \frac{N}{2}, \quad \text{отсюда} \quad Fk = \frac{2F_s}{N}.$$

Пример

Для *Fs* = 8 кГц, *N* = 256 получаем $Fk = 62{,}5$ Гц.

Для сигналов с бóльшими частотами дискретизации применяют окна бóльшего размера (см. табл. 2.1), при этом частота кадров остается приблизительно постоянной.

Другое соображение при выборе частоты кадров анализа заключается в следующем. Динамика временных огибающих спектра может быть охарактеризована *модуляционным спектром* (МС) – результатом спектрального анализа огибающих [5].

Для адекватного представления динамики огибающих частота следования кадров должна быть по крайней мере в два раза больше частотного диапазона модуляционного спектра.

Исследования показали, что максимум МС речевых сигналов находится в области 4 Гц, а основная часть МС находится в диапазоне до 20 Гц. Таким образом, для анализа РС достаточно частоты следования кадров 40 Гц.

Значения размеров и частоты кадров анализа РС для различных частот дискретизации сигнала приведены в таблице 2.2.

Таблица 2.2 — Размеры окон и частоты кадров анализа РС для различных частот дискретизации сигнала

Частота Fs, кГц	Размер окна и частоты кадров анализа		
	Размер кадра, $Nfft$	Шаг кадров, L	Частота кадров, Fk, Гц
8–10	256	128	32–40
20	512	256	40
40	1024	512	40
более 40	2048	1024	40

2.1.4 Широкополосная и узкополосная спектрограмма

Выбор частотного и временного разрешения КСА определяется особенностями анализируемого сигнала и задачами анализа. Разные параметры анализа приводят к разным представлениям свойств сигнала. В зависимости от длины окна спектрограмма может представить сигнал с различным временным и частотным разрешением.

Соотношение между временным и частотным разрешением определяется принципом неопределенности:

$$\Delta F = \frac{1}{\Delta T} = \frac{1}{NT} = \frac{F_s}{N}.$$

Средние частоты полос дискретного спектра при этом такие:

$$fn = 0, 1/\Delta T, 2/\Delta T, 3/\Delta T, \ldots = n \times \Delta F,$$

где ΔT – размер кадра;

ΔF – частотное разрешение.

Поскольку представление и анализ различных свойств РС требует применения окон различной длины, для их представления используют *узкополосную и широкополосную спектрограмму*.

Параметры широкополосной и узкополосной спектрограмм речевого сигнала приведены в таблице 2.3.

Таблица 2.3 — Параметры широко- и узкополосных спектрограмм РС

Параметр спектрограммы	Значение	
Тип спектрограммы	Широкополосная	Узкополосная
Размер кадров, мс	4	25
Ширина спектральных полос, Гц	250	40
Шаг анализа, мс	1	10
Наблюдаемые свойства РС	Импульсная и формантная структура РС	Тональная структура РС

На рисунке 2.3 показаны узкополосная и широкополосная спектрограммы фрагмента речевого сигнала. На узкополосной спектрограмме отображается тональная структура речи, на широкополосной спектрограмме – импульсная и формантная структура речи.

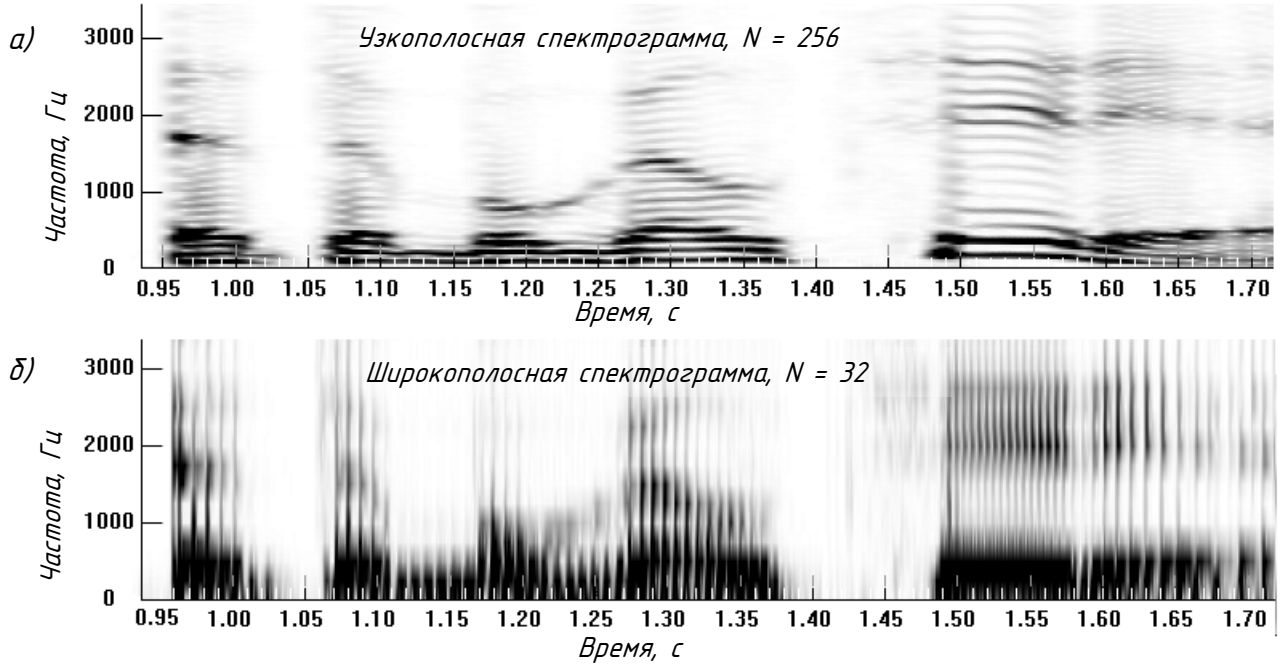


Рисунок 2.3 — Спектрограммы фрагмента речевого сигнала *Fs* = 8 кГц

2.2 Характеристики формы спектров

Характеристики формы спектров позволяют привести представление их свойств к небольшому числу параметров, тем самым уменьшить размерность их описания.

Характеристики формы спектров применяются для распознавания и детектирования различных акустических процессов, например, для распознавания типа шума, музыки и др. [6].

2.2.1 Средние по частоте характеристики

Основными параметрами спектров являются следующие:
– меры гладкости спектра (*spectral flatness measure, SFM*);
– меры тональности спектра (*spectral / frequency tonality measure, STM*);
– дисперсия спектра (*spectral variance, S_VAR*);
– энтропия спектра (*spectral entropy*);
– средняя амплитуда спектра (*spectral mean*);
– центр тяжести спектра (*spectral centroid*);
– ширина спектра (*spectral bandwidth*);
– минимальная и максимальная частота (*spectral roll-off point*);
– неравномерность спектра (*spectral roughness*);
– коэффициент тональности спектра (*harmonic coefficient*);
– спектральная асимметрия (*spectral skewness*);
– наклон спектра (*spectral slope*);
– куртозис (эксцесс) спектра (*spectral kurtosis*);
– степень концентрации спектра (*spectral spread*).

Подробное описание различных параметров спектров приведено в монографии [5]. Рассмотрим некоторые из них.

2.2.2 Меры гладкости спектра

Мера гладкости спектра (*spectral flatness measure*, *SFM*) определяется как отношение среднего (по частотам) геометрического и среднего арифметического значения на периодограммах, амплитудном спектре либо спектре мощности:

$$SFMx[k] = \exp < \ln\big(Ix(n,k)\big) >_F / < Ix(n,k) >_F,$$

$$SFMp[k] = \exp < \ln\big(Pxx(n,k)\big) >_F / < Pxx(n,k) >_F,$$

$$SFMa[k] = \exp < \ln(|Ax(n,k)|) >_F / < |Ax(n,k)| >_F \ ,$$

где $< \cdots >_F$ означает операцию усреднения по частоте дискретного спектра в выбранном частотном диапазоне.

Пример

$$< Ix(n,k) >_F = \frac{1}{Nf} \sum_{n=1}^{Nf} Ix[n,k] \ ,$$

Также используют логарифмические представления этих мер:
SFM_dB = 10 lg*(SFM)*
Свойства мер гладкости спектра:
SFM ≤ 1.
SFM (тональная речь, тон) ≪ 1, *SFM* (шум) ≈ 1.
SFMmin_dB ≪ 0 дБ, *SFMmax_dB* = 0 дБ.

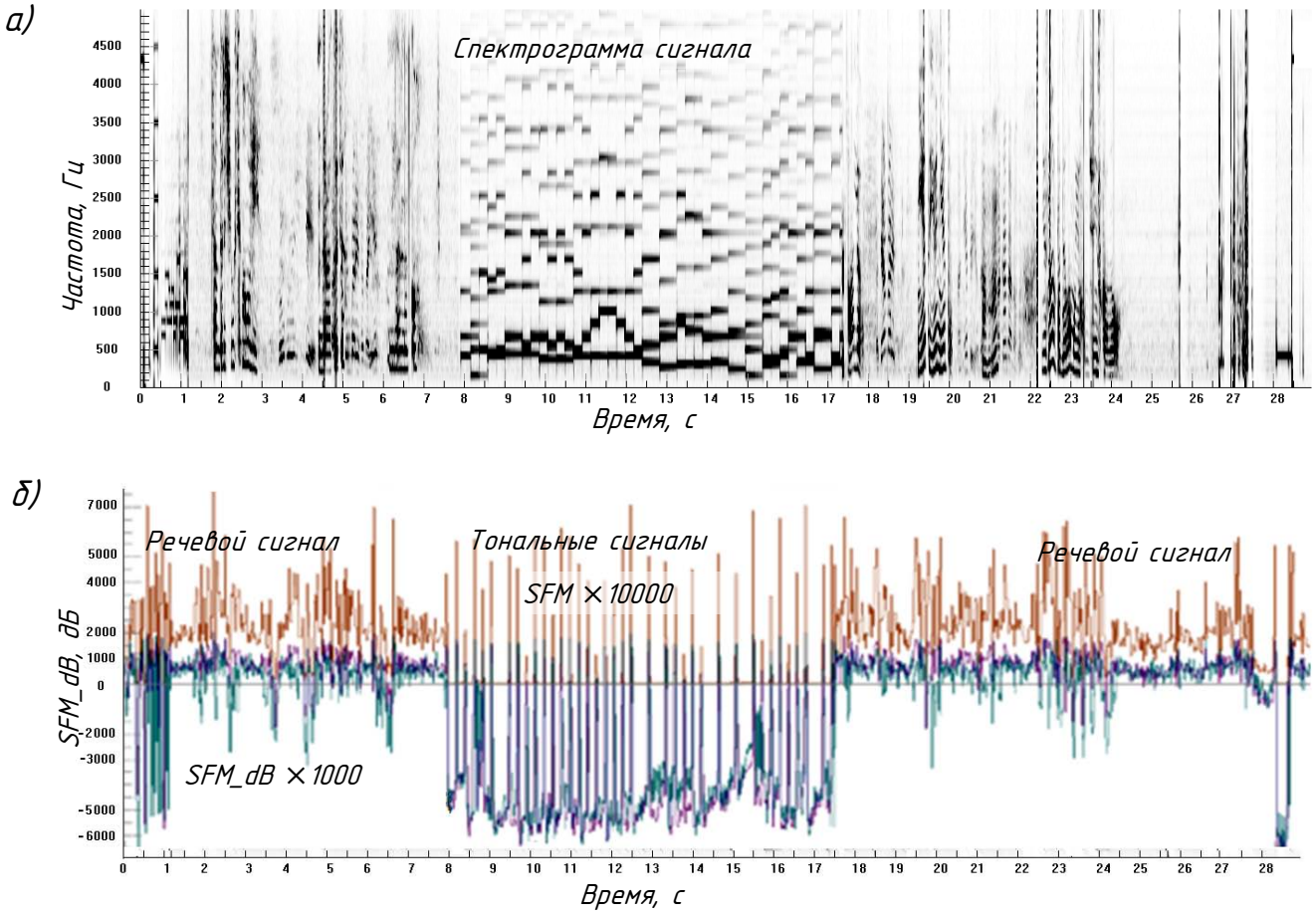


Рисунок 2.4 — Спектрограмма а) и меры гладкости б)
на фрагментах сигнала

2.2.3 Меры тональности спектра

Мера тональности спектра (*spectral tonality measure, STM*) определяется как величина, обратная мере гладкости спектра:

$$STM = \frac{1}{SFM}.$$

Также применяются логарифмические представления меры тональности:

$$STM\_dB = \min\{10\lg(STM), 60\text{дБ}\}$$

и нормализованная логарифмическая мера:

$$STMn\_dB = \min\{-SFM\_dB/60, 1\}.$$

Свойства мер тональности спектра:

STM (тональная речь, тон) $\gg 1$, *STM* (шум) ≥ 1;

STM_dB (тональная речь) > 40 дБ;

STMn_dB = 1 для тональных сигналов, *STM_dB* = 0 для шумов.

Пример применения меры гладкости для детектирования тональных сигналов телефонной линии приведен на рисунке 2.4.

Из рисунка следует, что *SFM* позволяет эффективно детектировать присутствие в сигнале тональных компонент.

2.2.4 Нормированная дисперсия амплитудного спектра

Нормированная дисперсия спектра (*energy-normalized spectral variance, S_VAR*) определяется следующим образом:

$$S\_VAR(k) = \frac{< [|X(n,k)| - < |X(n,k)| >_F]^2 >_F}{[< |X(n,k)|^2 >_F]}.$$

Чтобы мера характеризовала именно разброс спектральных амплитуд, а не крупномасштабную неравномерность спектра, спектр выравнивают, например, с помощью спектральной инверсной фильтрации.

Свойства нормированной дисперсии:

S_VAR ≤ 1;

S_VAR(шум) $\ll 1$;

S_VAR(тональная речь, тон) ≈ 1.

Бóльшая величина *S_VAR(k)* соответствует областям гармоник, малая величина соответствует областям доминирования шума.

Меры *SFM* и *S_VAR* являются взаимно обратными.

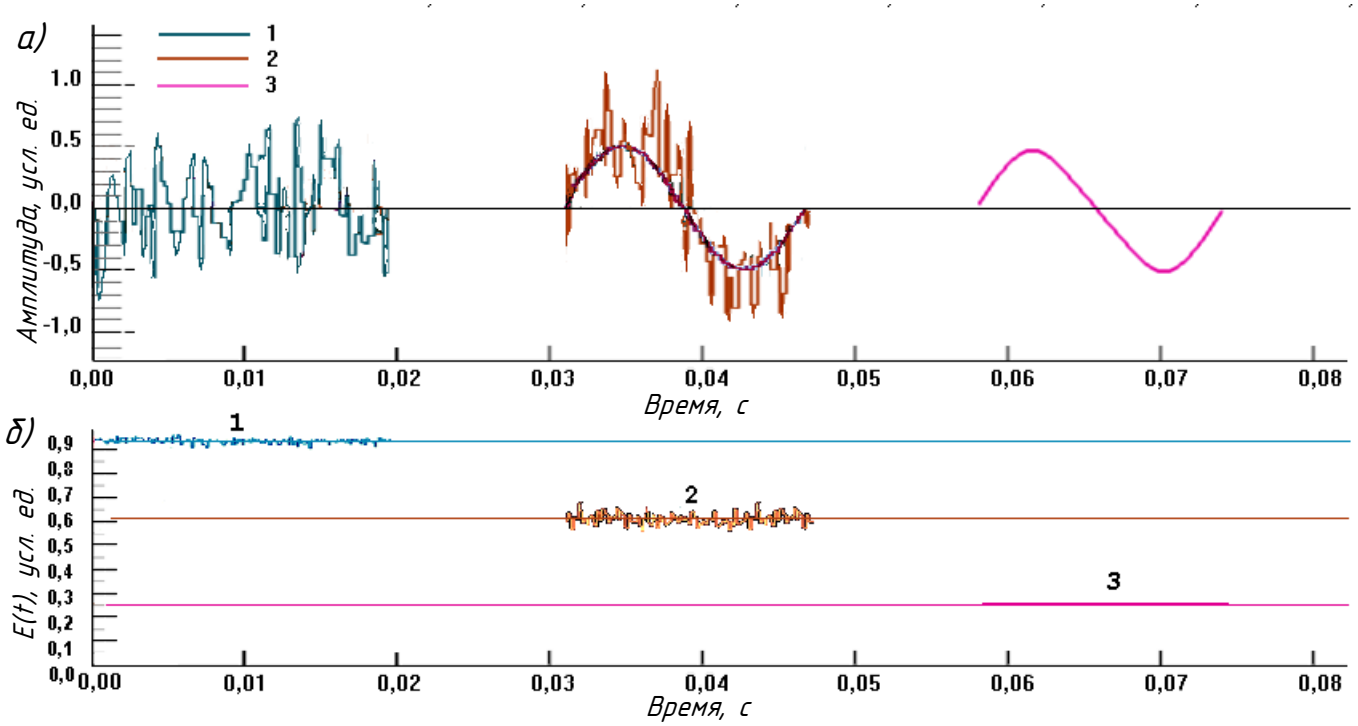
2.2.5 Энтропия спектра

Энтропия спектра определяется следующим образом:

$$E(k) = \frac{-1}{\ln(N)} \sum_{n=1}^{N} p_n(k) \ln\left(p_n(k)\right),$$

где «вероятность» p_n в полосе частот n определяется как нормализованная величина спектра мощности:

$$P_n(k) = \frac{|X(n,k)|^2}{\sum_{n=1}^{N} |X(n,k)|^2}.$$



а) – осциллограмма сигналов, б) – оценки энтропии спектра сигналов.
1 – белый шум; 2 – *SIN* + БШ (σ = 3000); 3 – *SIN* + БШ (σ = 3).

Рисунок 2.5 — Пример оценок энтропии спектра

Рассмотрим пример оценки энтропии для трех сигналов: белого шума $x1[i] = w[i]$ (СКО =3000), смеси тонального сигнала с шумом $x2[i] = \sin(\Omega\, i) + w[i]$ и смеси тонального сигнала с белым шумом малой амплитуды $x3[i] = \sin(\Omega\, i) + 0{,}01 \times w[i]$. Из рисунка видно, что энтропия, близкая к «1», соответствует белому шуму (БШ), а энтропия чистого тона (SIN) близка к величине «0».

2.3 Меры близости и подобия спектров

Назначение мер – оценка близости сигналов. Оценки близости сигналов во временной области чувствительны к синхронизации процессов, спектральные меры менее чувствительны, поскольку амплитудные спектры в некоторой мере инвариантны к сдвигу сигнала по времени. В этом параграфе рассмотрены *объективные меры*, не учитывающие свойства человеческого слуха. Меры, используемые для сравнения речевых сигналов, учитывающие особенности человеческого слуха, будут рассмотрены в следующей главе.

2.3.1 Дистанции между спектрами

Оценка дистанции между спектрами применяется для сравнения образцового спектра со спектром исследуемого сигнала. В случае, когда сигналы получены из одного источника, то есть нормированы одинаково, могут использоваться *разностные меры*. Для приведения дистанций к заданному диапазону (независимому от нормировки сигналов) дистанция нормализуются. В случае, когда сигналы получены из различных источников, необходимо учесть возможное различие нормировки сигналов и использовать инвариантные к отличию нормировки меры.

Детальное описание расстояний между спектрами приведено в работах [5, 7, 8].

Спектры можно рассматривать как векторы, поэтому меры дистанции спек-

тров могут быть определены аналогично мерам дистанции векторов. Меры дистанции могут применяться как к мгновенным амплитудным спектрам $Ax(n,k) = |X(n,k)|$, периодограммам $Ix(n,k) = |X(n,k)|^2$), так и к средним спектрам $Ax(n), Pxx(n)$. В дальнейшем для краткости будем их обозначать как $Xn(k)$ или, там, где это не мешает пониманию смысла, опуская временной индекс.

Последовательность компонент спектра может быть представлена как вектор $\mathbf{X}= [X_1, X_2, \dots X_n, \dots X_N]^T$.

Компоненты векторов могут формироваться из спектров в некотором частотном диапазоне, отличном от полного частотного диапазона сигнала, и в выбранных частотных шкалах (не обязательно равномерной шкале ДПФ). С учетом этого введем следующие обозначения.

Операторы суммирования по времени и частоте:

$\sum_n Xn(k)$ *означает* $\sum_{n=n1}^{n=n2} X(n,k)$, $\sum_k Xn(k)$ *означает* $\sum_{k=k1}^{k=k2} X(n,k)$.

Оператор усреднения по времени: $< Xn >_T = \frac{1}{K}\sum_k X(n,k)$.

Оператор усреднения по частотам дискретного спектра:

$$< Xn >_F = \frac{1}{Nf}\sum_n X(n,k).$$

Оператор усреднения по времени и частотам:

$< X(n,k) >_{FT} = \frac{1}{Nf\ K}\sum_k\sum_n X(n,k))$,

где $n1$, $n2$ – индексы нижней и верхней границ выбранного частотного диапазона спектра ($Nf = n2 - n1 + 1$); K – количество кадров ($K = k2 - k1 + 1$).

В дальнейшем используются следующие векторные обозначения:

вектор-столбец: $\mathbf{X} = [X_1, X_2, \dots X_N]^{\mathrm{T}}$, $[\ ]^{\mathrm{T}}$ – операция транспонирования;

вектор-строка: $\mathbf{X} = [X_1, X_2, \dots X_N]$.

Длина вектора (евклидова норма в R^n): $\|\boldsymbol{X}\| = (\mathbf{X}^T\mathbf{X})^{\frac{1}{2}} = [\sum_n Xn^2]^{\frac{1}{2}}$

Скалярное произведение векторов: $\mathbf{X}^{\mathrm{T}}\ \mathbf{Y}$ = <**X**, **Y**> = $\sum_n Xn\ Yn$

Меры различия спектров

Мера различия спектров (*spectral discrepancy, SD*) [5]:

$$SD(\mathbf{X},\mathbf{Y}) = \min\{1, \sum_n |Xn - Yn| / \sum_n Yn\}.$$

Логарифмическая мера различия (*logarithmic spectral distance, LSD*) [9]:

$$LSD(\mathbf{X},\mathbf{Y}) = \left[\frac{1}{Nf}\sum_n [20\lg|Xn| - 20\lg|Yn|]^2\right]^{1/2}.$$

Приведенные меры предполагает одинаковую нормировку сигналов при этом они не зависят от общей (одинаковой для обоих сигналов) нормировки.

Дистанции в других пространствах

Поскольку спектры могут вычисляться через коэффициенты линейного предсказания (КЛП), а кепстры могут быть вычислены через спектры, то в ряде случаев дистанции между спектрами могут быть представлены в пространстве КЛП или в пространстве кепстров.

Дистанция в пространстве кепстров может также быть определена следующим соотношением:

$$Dc(\mathbf{X},\mathbf{Y}) = \sum_{m=1,M} |CCx(m) - CCy(m)|.$$

При определенных условиях мера *LSD* может быть выражена через коэффициенты кепстров [9]:

$$LSD(\mathbf{X},\mathbf{Y}) = 20\lg(e)\left[2\sum_{m=1,M}[CCx(m) - CCy(m)]^2\right]^{1/2}.$$

Здесь не учитываются нулевые коэффициенты кепстров, содержащие информацию о средней амплитуде спектров, поэтому данные меры не зависят от уровня нормировки сигналов. Другие меры на основе коэффициентов кепстра рассмотрены в книге [10].

Дистанция в пространстве коэффициентов линейного предсказания

Поскольку спектры могут вычисляться через коэффициенты линейного предсказания (КЛП), то дистанция между спектрами может быть выражена через КЛП. *Оценка максимального правдоподобия* (*log-likelihood ratio, LLR*) вычисляется через вектор КЛП чистой и зашумленной речи [11]:

$$dLLR(\mathbf{A}_d, \mathbf{A}_c) = \log(\mathbf{A}_d^{\mathrm{T}}\mathbf{R}_c\mathbf{A}_d / \mathbf{A}_c^{\mathrm{T}}\mathbf{R}_c\mathbf{A}_c),$$

где $\mathbf{A}_d$ – вектор КЛП искаженной речи;

$\mathbf{A}_c$ – вектор КЛП чистой речи;

$\mathbf{R}_c$ – автокорреляционная матрица чистой речи.

2.3.2 Меры подобия спектров

В случае, когда источники сигналов разные, то есть сигналы могут быть нормированы неодинаково, разностные меры неприменимы. В этих случаях применяют меры, основанные на мерах близости (подобия) спектров. Рассмотрим две из них.

Косинусное расстояние между векторами

Косинусная метрика сходства спектров определяется как угол между векторами:

$$\cos\varphi = \mathbf{X}^{\mathrm{T}}\mathbf{Y}/(||\mathbf{X}||\,||\mathbf{Y}||) = \langle\mathbf{X},\mathbf{Y}\rangle/(||\mathbf{X}||\,||\mathbf{Y}||)$$

На основе нее определяется *косинусное расстояние* [8]

$$d(\mathbf{X},\mathbf{Y}) = 1 - \cos\varphi$$

Корреляционное расстояние Пирсона

Коэффициент корреляции Пирсона определяется следующим соотношением

$$\rho(\mathbf{X},\mathbf{Y}) = \frac{\sum_n (Xn - \langle Xn\rangle)(Yn - \langle Yn\rangle)}{[\sum_n (Xn - \langle Xn\rangle)^2 \sum_n (Yn - \langle Yn\rangle)^2]^{1/2}}$$

На основе него определяется корреляционное расстояние Пирсона

$$d(\mathbf{X},\mathbf{Y}) = 1 - \rho(\mathbf{X},\mathbf{Y})$$

Важно, что эти же меры можно применить как к спектрограммам, так и к последовательностям амплитуд разделенных по полосам сигналов (после гребенки фильтров).

2.4 Меры кратковременной динамики спектров

2.4.1 Описание мер динамики спектров

Меры кратковременной динамики определяются на основе сравнения спектров на соседних кадрах, обычно как первые или вторые производные по времени. Меры динамики спектров характеризуют временные связи (подобие или различие) спектров текущего и предшествующего кадра. Эти простейшие меры характеризуют изменение аудиоинформации от кадра к кадру.

Дистанция изменения спектра (*spectral derivative distance*, *SDD*)

$$SDD[k] = \frac{< (|X(n,k)| - |X(n,k-1)|)^2 >_F}{[< X(n,k)^2 >_F < X(n,k-1)^2 >_F]^{1/2}}.$$

Меры роста и уменьшения спектра

$$SDR\_up[k] = \sum_n HR\{|X(n,k)| - |X(n,k-1)|\}.$$

$$SDR\_dn[k] = \sum_n HR\{|X(n,k-1)| - |X(n,k)|\},$$

где $HR\{\}$ – функция полуволнового выпрямления (*half-wave rectification*, *HR*): $HR\{x\} = \frac{1}{2}(x + |x|)$.

Дистанция логарифмов спектров (*spectral distance logarithms*, *SDL*)

$$SDL(k) = \frac{1}{Nf}\sum_n \ln\left(\frac{|X(n,k)|}{|X(n,k-1)|}\right).$$

Корреляция спектров (*spectral correlation*, *SC*)

Средняя по частоте корреляция между спектрами соседних кадров [5]:

$$SC(k) = \frac{< |X(n,k)||X(n,k-1)| >_F}{[< |X(n,k)^2| >_F < |X(n,k-1)|^2 >_F]^{1/2}}.$$

На основе корреляции определяется *средняя дистанция* (расстояние) между спектрами соседних кадров:

$$SD(k) = 1 - SC(k).$$

Свойства нормализованных мер:

– инвариантность к амплитуде сигнала;

– основной вклад вносят мощные спектральные компоненты сигнала;

– на тональных сигналах величина близка к единице.

Спектральный поток (*spectral flux*, *SF*)

Спектральный поток является мерой изменения формы спектра, описывающей дистанцию его обновления между двумя соседними кадрами. Спектральный поток позволяет детектировать текущие изменения спектра процесса. Предложено несколько мер (нормированных и ненормированных) спектрального потока. Приведем некоторые из *нормированных мер*.

В книге [5] спектральный поток определяется как косинусная дистанция между спектрами соседних кадров:

$$SFx(k) = 1 - SC(k).$$

В работе [6] спектральный поток определяется через спектры мощности:

$$SFp(k) = 1 - \frac{< Pxx(n,k)Pxx(n,k-1) >_F}{[< Pxx(n,k)^2 >_F < Pxx(n,k-1)^2 >_F]^{1/2}}.$$

Определение спектрального потока через логарифмы модулей спектров дано в [6]:

$$SF_L(k) = < |\log|X(n,k)| - |\log|X(n,k-1)| >_F.$$

Свойства нормализованных мер спектрального потока:

– инвариантность к амплитуде сигнала;

– основной вклад вносят мощные спектральные компоненты сигнала;

– на тональных сигналах величина близка к единице.

2.4.2 Применение мер динамики спектров

Меры динамики спектров нашли применение в решении многих задач:
– детектирование различных типов звуков;
– детектирование речевой активности (*voice activity detector, VAD*);
– детектирование пауз (*silence activity detector, SAD*);
– детектирование фронтов сигналов (*onset/offset detection*).

На рисунке 2.6 приведены графики критериев *SF*(*t*), *SDD*(*t*) на тестовом сигнале длительностью 360 с, состоящем из последовательности различных типов звуков: тональная музыка (20–100 с) → стационарный «розовый шум» (100–190 с) → речь (190–240 с) → речеподобный модулированный шум (240–360 с).

Из рисунка видно, что обе меры динамики спектра ведут себя по-разному, позволяя при этом различать сигналы с разными свойствами.

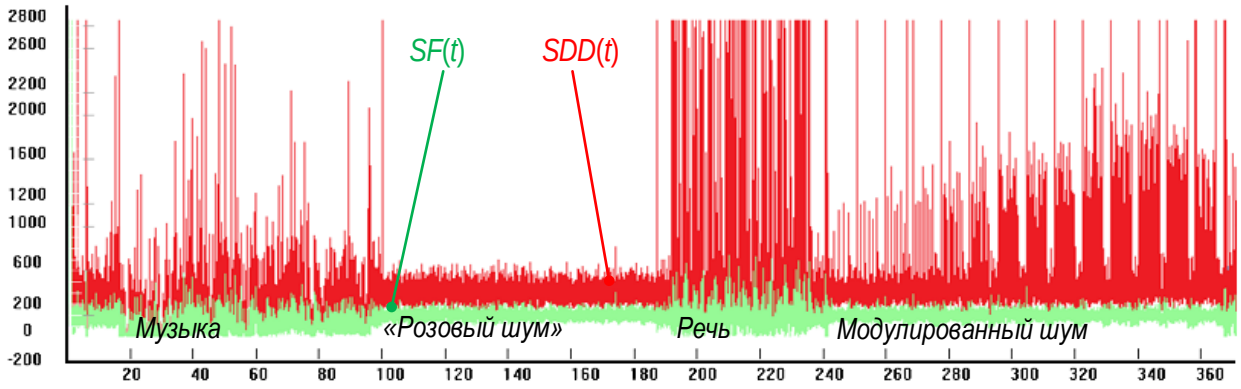


Рисунок 2.6 — Меры динамики спектров на тестовом сигнале

2.5 Частотно-временные характеристики спектров

Огибающие спектров в каждой частотной полосе (*spectral envelopes*) могут рассматриваться как сигналы, дискретизированные с частотой следования кадров. Это позволяет взглянуть на них с новой стороны и применить к ним характеристики, используемые при анализе сигналов. В данном параграфе рассмотрены некоторые из таких характеристик.

2.5.1 Средние по времени характеристики огибающих

Множество значений амплитуд спектра на кадрах можно охарактеризовать в каждой полосе частот функцией распределения амплитуд.

Функция распределения амплитуд может быть представлена средним значением, дисперсией, куртозисом и другими характеристиками.

Эти параметры содержат важную информацию о процессе и применяются в решении ряда практических задач. Однако обычно ограничиваются средними значениями.

Средние спектры определяются следующими формулами.

Средний амплитудный спектр:

$$Ax(n) =< |X(n,k)| >_T .$$

Средний спектр мощности:

$$Pxx(n) =< |X(n,k)|^2 >_T .$$

Средний геометрический спектр:

$$Gx(n) = \left[\prod_{k=1}^{K} |X(n,k)|\right]^{1/K} = \exp[< \ln|< X(n,k)| >_T] ,$$

где K – число кадров.

Усреднение может выполняться по всему файлу либо по скользящим фрагментам. В случае усреднения по фрагментам характеристики представляют собой результаты низкочастотной фильтрации огибающих и сами меняются во времени. Это же относится к другим характеристикам.

2.5.2 Средняя по времени корреляция между спектрами соседних кадров

Средняя по времени корреляция между комплексными спектрами соседних кадров задается следующим выражением:

$$SC(n) = \frac{< |X(n,k)^* X(n,k-1)| >_T}{[< X(n,k)|^2 >_T < X(n,k-1)|^2 >_T]^{1/2}} ,$$

где * – символ комплексного сопряжения.

На её основе определяется средняя по времени дистанция между спектрами соседних кадров:

$$SD(n) = 1 - SC(n) .$$

2.5.3 Дистанция между спектрограммами

Еще одной важной частотно-временной характеристикой спектров является дистанция между спектрограммами и степень подобия спектрограмм.

Рассмотрим спектрограммы двух сигналов $x[n]$, $y[n]$, для которых построены спектрограммы амплитудных спектров $|X(n, k)|$, $|Y(n, k)|$ или спектрограммы спектров мощности $Pxx(n, k)$, $PYY(n, k)$.

Для фрагментов спектрограмм дистанции по массиву значений частот спектров $1 \le n \le N$ и массиву кадров вокруг кадра с номером m: $-K \le k \le m + K$ могут быть определены дистанции и меры подобия.

Линейные дистанции

$$D_P(x,y) = < |Pxx(n,k) - Pyy(n,k)| >_{FT} ,$$

$$D_A(x,y) = < \big||X(n,k)| - |Y(n,k)|\big| >_{FT} .$$

Логарифмические дистанции

$$LD_P(x,y) = < 10 \lg \left[\frac{P(n,k)}{P(n,k)}\right] >_{\_FT}.$$

$$LD_A(x,y) = < 20 \lg \left[\frac{|X(n,k)|}{|Y(n,k)|}\right] >_{\_FT}.$$

Линейная и логарифмическая дистанции не являются инвариантными к масштабу сигналов $x(t)$, $y(t)$ мерами.

То есть, если сигналы взяты из разных источников, то для одинаковых сигналов, взятых с разным коэффициентом усиления, дистанция не будет равна нулю: $D(x,y) \neq D(x,gy), D(x,gx) \neq 0$.

Для устранения данной проблемы можно использовать *меры подобия спектрограмм*.

Косинусные меры подобия являются нормализованным мерами, инвариантными к масштабу сигналов *x*(*t*), *y*(*t*). Косинусные меры задаются следующими соотношениями:

$$C_p(x,y) = \frac{< Pxx(n,k)Pyy(n,k) >_{FT}}{[< Pxx(n,k)^2 >_{FT} < Pyy(n,k)^2 >_{FT}]^{1/2}},$$

$$C_A(x,y) = \frac{< |X(n,k)||Y(n,k)| >_{FT}}{[< |X(n,k)|^2 >_{FT} < |Y(n,k)|^2 >_{FT}]^{1/2}}.$$

Тогда

$$C(x,y) = C(x,gy).$$

Нормализованная дистанция между спектрограммами определяется через меру подобия следующим образом:

$$D(x,y) = 1 - C(x,y).$$

Обобщением этих мер является коэффициент корреляции Пирсона.

2.5.4 Коэффициент корреляции Пирсона для спектрограмм

Коэффициент корреляции Пирсона (*Pearson's correlation coefficient*, *PCC*) является нормализованной мерой подобия, инвариантной к масштабу сигналов *x*(*t*), *y*(*t*).

Для спектрограмм коэффициент корреляции вычисляется по массиву значений частот спектров $1 \le n \le N$ и массиву кадров вокруг кадра с номером m: $-K \le k \le m + K$. Мера задается следующими соотношениями:

$$PCC_{xy} = \frac{V_{xy} - \mu x \cdot \mu y}{\left[(V_{xx} - \mu x^2) \cdot (V_{yy} - \mu y^2)\right]^{1/2}},$$

где $\mu x = \frac{1}{KN}\sum_k \sum_n |X(n,k)|$,

$\mu y = \frac{1}{KN}\sum_k \sum_n |Y(n,k)|$;

n – частотные компоненты: $1 \le n \le N$;

$2K + 1$ кадров вокруг целевого кадра m: $m - K \le k \le m + K$, всего учитывается $KN = N \times (2K + 1)$ компонент спектрограммы;

$V_{xx} = \frac{1}{KN}\sum_k \sum_n |X(n,k)|^2$;

$V_{yy} = \frac{1}{KN}\sum_k \sum_n |Y(n,k)|^2$;

$V_{xy} = \frac{1}{KN}\sum_k \sum_n |X(n,k)| \times |Y(n,k)|$.

Нормализованная дистанция между спектрограммами определяется через коэффициент Пирсона следующим образом:

$$D_{xy} = 1 - PCC_{xy}.$$

Коэффициент корреляции Пирсона является эффективным критерием, например, для поиска фрагмента аудиозаписи в базе данных по образцу [12].

2.5.5 Модуляционный спектр

Модуляционный спектр (МС) позволяет выделить информацию о периодических модуляциях амплитуды сигнала. Обычный спектральный анализ не позволяет обнаружить периодическую структуру огибающей сигнала, если сигнал не является когерентным.

Смысл модуляционного спектрального анализа (МСА) заключается в применении спектрального анализа к временным огибающим амплитуд спектра сигнала.

Это утверждение иллюстрирует следующий пример. Сигнал с частотой дискретизации *Fs* = 11025 Гц представляет собой случайный (слабо коррелированный) шум с речеподобным спектром. Шум модулирован во времени по амплитуде с частотой 12,5 Гц. Задача заключается в обнаружении периодической модуляции сигнала. Спектральный анализ проводился с окном 8192 отсчетов, что обеспечивает спектральное разрешение 1,35 Гц и обнаружение компоненты с частотой 12,5 Гц.

На рисунке 2.7 а) показана осциллограмма сигнала и его огибающая мощности в окне 20 мс, также показано окно, в котором вычислялся спектр сигнала и спектр огибающей.

На рисунке 2.7 б) показаны спектры сигнала и его огибающей мощности в децибелах в диапазоне нижних частот.

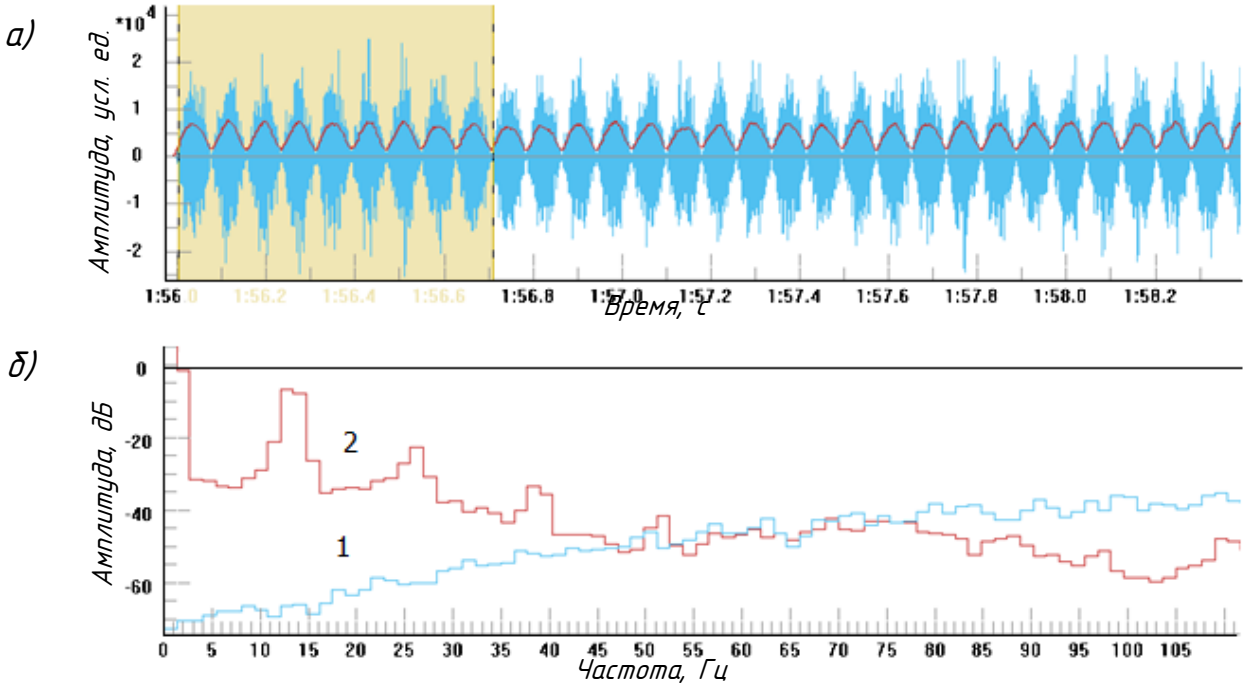


Рисунок 2.7 — Осциллограмма сигнала: а) окно анализа и огибающая мощности, б) спектры (дБ) (1): сигнала, (2): огибающей мощности

Из рисунка 2.7 видно, что спектр самого сигнала является относительно плоским, в то время как спектр огибающей имеет ярко выраженный пик на частоте 12,5 Гц.

Таким образом, спектральный анализ огибающих мощности позволяет обнаруживать периодическую структуру некогерентного сигнала (например, работающих механизмов и др.).

Модуляционный спектральный анализ является обобщением рассмотренного примера. Спектральный анализ выполняется для огибающих мощности в различных

частотных диапазонах сигнала, то есть для временной последовательности амплитуд спектров.

Разделение сигнала по частотным полосам может быть выполнено с помощью гребенки фильтров или ДПФ. После этого вычисляется последовательность значений огибающих спектров мощности. Огибающие берутся с частотой следования кадров *Fk* (прореживаются по времени). Предварительно огибающие подвергаются низкочастотной фильтрации для того, чтобы ограничить полосу частот и избежать эффекта наложения спектров. Полоса фильтров низкой частоты (ФНЧ) должна быть не более половины частоты следования кадров.

Далее к огибающим применяется спектральный анализ с использованием дискретного преобразования Фурье или гребенки фильтров.

Результатом такой операции является *модуляционный спектр* (*modulation spectrum, MS*). Вычисление модуляционного спектра для огибающих амплитуд спектров $|X(f, k)|$, $k = 1, 2, \ldots$определяется следующими соотношениями:
если СА сигнала выполняется с использованием ДПФ

$$MSx(f, Fm) = |DFT\{|X(f, k)|\}|^2 ,$$

если СА сигнала выполняется с использованием гребёнки ФНЧ

$$MSx(f, Fm) = \left|DFT\{\text{ФНЧ}\{Filt(f)\{x[n]\}\}\}\right|^2 ,$$

где Fm – модуляционные частоты;

f – полоса частот временной огибающей спектра мощности.

На рисунке 2.8 приведена схема вычисления модуляционного спектра с использованием банка фильтров.

Рисунок 2.8 — Схема вычисления модуляционного спектра [13]

На вход поступает анализируемый сигнал. Сигнал разделяется в банке фильтров на полосовые сигналы. Далее в каждой полосе частот вычисляется огибающая сигналов. Для дальнейшего спектрального анализа, выполняемого от кадра к кадру с частотой следования кадров (*Fk = Fs/100*), частота полосы огибающей предварительно ограничивается величиной 28 Гц, чтобы избежать эффекта наложения спектров. После этого выполняется нормализация амплитуд текущих огибающих, к кадру каждого из которых применяется дискретное преобразование Фурье, после чего вычисляется дискретный спектр мощности.

Матрица значений спектров мощности в частотных полосах сигнала образует матрицу значений модуляционного спектра $MSx(f, Fm)$.

Пример представления модуляционного спектра приведен на рисунке 2.9.

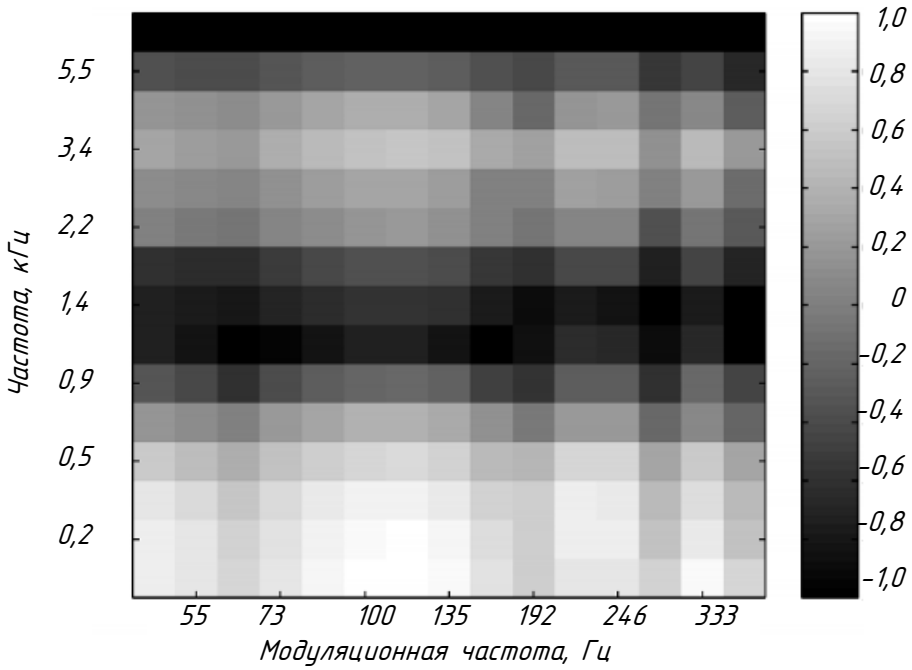


Рисунок 2.9 — Пример представления модуляционного спектра [14]

В таблице 2.4 перечислены акустические процессы в различных диапазонах модуляционных частот модуляционного спектра.

Таблица 2.4 — Акустические процессы в различных диапазонах МС

Диапазон модуляционных частот, Гц	Акустический процесс
0-1	Стационарный шум
1-2	Реверберация
2-4-8-16-32	Модуляция звуков речи
32-100	Периодические импульсные события
более 100	Импульсы основного тона РС

Различные инструменты КСА являются эффективным средством представления свойств акустических периодических процессов различной природы: речевых сигналов, тональных сигналов, музыки, шумов механизмов и многих других. Рассмотренные методы составляют лишь малую часть инструментария КСА.

В заключение отметим важный момент – тесную связь спектрального анализа и моделирования сигналов [14]. Практически невозможно выполнить спектральный анализ с высоким качеством без формулировки дополнительных гипотез. Данная группа методов спектрального анализа, получившая название параметрического спектрального анализа, подробно освещена в книге Марпла [15].

Литература

1. Boashash B. (Ed.). Time-Frequency Signal Analysis and Processing. A Comprehen-sive Reference. Academic Press, 2015, 1056 pp.

2. Иванушкин С. А. и др. Основные спектральные характеристики речевых сигналов и их определение // Компьютерное моделирование в фундаментальных и прикладных исследованиях, 2021.

3. Рабинер Л., Гоулд Б. Теория и применение цифровой обработки сигналов // М.: Мир, 1978. — 848 с. (п. 2.25. Секционированные свертки).

4. Отнес Р., Эноксон Л. Прикладной анализ временных рядов. Основные методы // М.: Мир, 1982. — 432 с.

5. Lerch A. An introduction to audio content analysis. Applications in signal processing and music informatics // John Wiley & Sons, Inc., 2012. 272 P

6. Leman A., Faure J., Parizet E. A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises // INTERSPEECH-2009, pp.1139–1141, 2009.

7. Cees H. Taal et al. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech – IEEE Transactions on audio, speech, and language processing – vol 19, № 7 – 2011 – URL: https://ieeexplore.ieee.org/document/ 5713237

8. Шейнман Е. Л. Распознавание объектов и определение дистанции по эталонам базы данных спектров сигналов // Фундаментальная и прикладная гидрофизика, 2019. Т. 12, №2. С.20–26.

9. Vary P., Martin R. Digital speech transmission: enhancement, Coding and error concealment, Wiley, 2006, 644 P.

10. Тампель И.Б., Карпов А.А. Автоматическое распознавание речи. Учебное пособие. – СПб.: Университет ИТМО, 2016. – 138 с.

11. Маркел Дж. Д., Грей А. Х. Линейное предсказание речи. – М.: Связь, 1980. – 308 с.

12. Xiao B, et al. Overlapped speech detection using long-term spectro-temporal similarity in stereo recording // Proc. ICASSP 2011, pp.2516-2519.

13. Шарий Т.В., Черкашин Е.И. Распознавание голосовых команд управления роботом на основе спектра модуляции речевого сигнала // Вестник Донецкого национального университета, серия Г, Технические науки, № 2, 2018, С.41-48.

14. Макс Ж. Методы и техника обработки сигналов при физических измерениях. В 2-х томах — М.: Мир, 1983 (том 2).

15. Марпл С.Л. Цифровой спектральный анализ и его применение. – М.: Мир, 1990. – 584 с.

Вопросы и задачи

1. Как измерить (рассчитать) среднюю разницу фаз $<e^{j\Phi(f)}>$ двух сигналов $x[i]$, $y[i]$?

2. Докажите эквивалентность двух выражений среднего геометрического спектра: $G(n) = [\prod_{k=1,K} |X(n, k)|]^{1/K} = \exp[1/K \sum_{k=1,K} \ln|X(n, k)|]$.

3. Как получить объективную оценку близости формы двух дискретных спектров, не зависящую от нормировки сигналов?

4. Частота дискретизации сигнала 16 кГц. Спектры вычисляются с размером кадра 512, шагом 256 отсчетов. Какое число спектров усредняются на интервале 1 секунда?

5. Частота дискретизации сигнала 16 кГц. Спектры вычисляются с размером кадра 512, шагом 256 отсчетов. Какова частота Найквиста для модуляционного спектра?

6. Задана евклидова метрика вектора $||X|| = [\sum_n Xn^2]^{1/2}$. Докажите, что $||X - Y||^2 = \sum_n (Xn - Yn)^w$.

Глава 3
Кепстральный анализ сигналов

3.1 Определение кепстра

3.1.1 Терминология

Термин кепстр (*cepstrum*) был первоначально предложен перефразированием слова *spectrum* в статье Boget, B.P., Healy, M.J.R., Tukey, J.W., “The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-cepstrum and Saphe Cracking”, in Proceedings of Symposium on Time Series Analysis by Rosenblatt, M., 1963, pp.209–243.

Наряду с термином *cepstrum* был предложен ряд терминов, полученных аналогичным способом. Основанием явилось то, что кепстр является в некотором смысле спектром спектра. Наиболее распространёнными терминами, встречающимися в литературе до настоящего времени, являются следующие. Термин *quefrency*, закрепленный за X-осью кепстра, идентичный по смыслу задержке *y* автокорреляционной функции. Другой термин *rahmonics*, идентичный в гармоникам (*harmonics*) в спектре.

Первоначальное определение кепстра было таким [1]:

$$CCx(\tau) = FT^{-1}\{ Pxx(f) \},$$

где *Pxx*(*f*) – спектр мощности (спектральная плотность мощности), вычисляемый как усредненная по времени периодограмма; *FT* – преобразование Фурье.

3.1.2 Оценки кепстра

Кепстр является характеристикой производной от спектра мощности. Представим спектр как кадр сигнала (дискретного или непрерывного). Вычисляя обратное преобразование Фурье от спектра мощности, получим автоковариационную функцию:

$$Cxx(\tau) = FT^{-1}\{ Pxx(f) \},$$
$$Cxx[m] = DFT^{-1}\{ Pxx[n] \},$$

где *Pxx*(*n*) – дискретный спектр мощности;

n = 0, 1,…N-1 – размер спектра;

m = 0, 1,…N -1 – размер АКФ.

Кепстр вычисляется аналогичным образом, однако вместо спектра мощности берется его логарифм:

$$CCx[m] = DFT^{-1}\{\ln Pxx[n]\} = CEPS\{ Pxx[n]\}.$$

Для вычисления как АКФ, так и кепстра необходим полный спектр с положительными и отрицательными частотами. Поскольку *Pxx*(*f*) = *Pxx*(*-f*), то есть массив отсчетов спектра является симметричным, то в результате обратного преобразования Фурье получаем вещественную функцию кепстра.

Спектр мощности связан с кепстром соотношением:

$$Pxx[n] = \exp[DFT\{CCx[m]\}] = ICEPS\{CCx[m]\}$$

Для вычисления кепстра также необходим полный спектр мощности с положительными и отрицательными частотами. При переходе от кепстра к спектру мощности также необходимо использовать симметричный массив коэффициентов кепстра:

$$Pxx[N-n] = Pxx[n], n = 1,2,..N/2,$$
$$CCx[N-n] = CCx[n], n = 1,2,..N/2.$$

Реальный кепстр также определяют через амплитудный спектр $Ax[n]$:

$$CCx[m] = DFT^{-1}\{\ln Ax[n]\}.$$

На кадрах сигнала реальный кепстр определяется следующим соотношением:

$$CCx(m,k) = IDFT\{\ln |X[n,k])|\}.$$

Комплексный кепстр определяется через комплексный спектр:

$$X[n] = |X[n]|\exp(j\,\Phi x[n]),$$
$$CCx[m] = DFT^{-1}\{\ln X[n]\} = DFT^{-1}\{\ln |X[n]|\} + j\,DFT^{-1}\{\Phi x[n]\}.$$

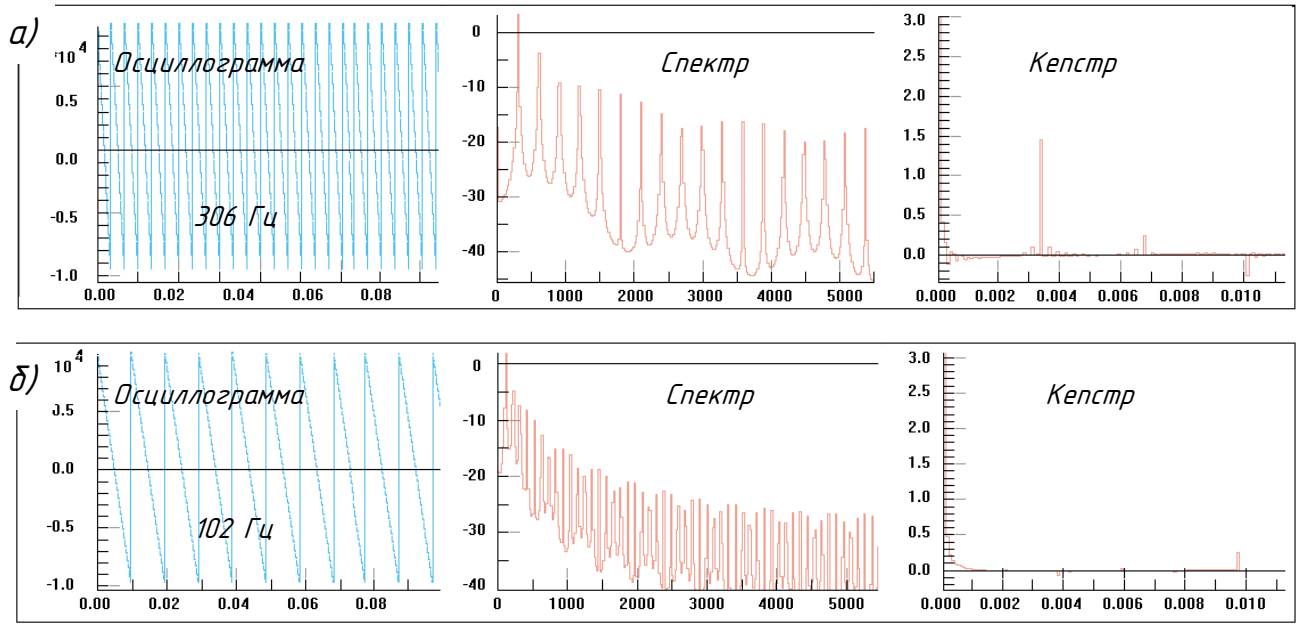
Если сигнал моделируется авторегрессионным процессом, а его спектр вычисляется по коэффициентам линейного предсказания, тогда кепстр может быть вычислены по спектру линейного предсказания либо непосредственно по КЛП [2].

3.1.3 Свойства кепстра

Важнейшее свойство кепстров заключается в том, что эффекты источника сигнала и канала распространения в пространстве кепстров могут быть разделены. Следующие преобразования доказывают это:

$$y(t) = h * s(t),$$
$$Y(f) = S(f)H(f),$$
$$ln\,Y(f) = \ln S(f) + \ln H(f),$$
$$FT\{\ln |Y(f)|\} = FT\{\ln |S(f)|\} + FT\{\ln |H(f)|\},$$
$$CCy = CCs + CCh\,.$$

Другим полезным свойством является то, что подобно тому, как в спектре тонального сигнала на частоте тона формируется пик, в кепстре периодического импульсного сигнала (звуков речи, звуков работающих механизмов) формируется пик с частотой, соответствующей частоте следования импульсов.



а) частота периода 306 Гц), б) частота периода 102 Гц)

Рисунок 3.1 — Осциллограммы периодических импульсных сигналов, логарифм среднего спектра сигнал, средний кепстр сигнала

Данные соотношения иллюстрируются на рисунках 3.1 а) и б). На каждом из рисунков представлены осциллограмма периодического сигнала, его средний спектр и средний кепстр.

Периодическим импульсным сигналам с периодом To соответствует периодический спектр с шагом следования пиков, равным частоте следования импульсов $Fo = 1/To$, которому соответствует средний кепстр с пиком, соответствующим задержке To. Таким образом, кепстр позволяет определить как период, так и частоту следования импульсов.

Еще одним важным свойством кепстра является независимость его коэффициентов от уровня сигнала. Изменение амплитуды сигнала приводит добавлению к логарифму спектра константы, которая лишь изменяет значение нулевого коэффициента кепстра, при этом другие коэффициенты не меняются.

3.2 Представление кепстра в различных шкалах

В зависимости от выбора шкалы аргумента кепстр может быть представлен тремя способами:

– в шкале номеров коэффициентов m;

– в шкале временных задержек;

– в шкале частот.

Рассмотрим соотношение между ними.

Соотношение между шкалой номеров и шкалой временных задержек такое же, как у автокорреляционной функции:

$$CCx[m] = CCx(T \times m) = CCx(Tm) = CCx\left(\frac{m}{F_S}\right).$$

Шкала частот однозначно связана со шкалой задержек и номеров коэффициентов:

$$CCx[m] = CCx(Tm) = CCx\left(\frac{1}{Fm}\right) = CCx\left(\frac{m}{Fs}\right).$$

То есть

$$Fm = Fs/m = 1/Tm.$$

Поскольку кепстр для спектра с окном *N* содержит *N/2* коэффициентов, то минимальная наблюдаемая частота равна: $Fmin = 2Fs/N \Leftrightarrow Nmin = 2Fs/Fo$

Таким образом, длина окна кепстра ограничивает минимальную отображаемую частоту основного тона.

Шкалы номеров и задержек являются линейными, а шкала частот — нелинейной. Шаг по частоте увеличивается с уменьшением номера коэффициента и с увеличением частоты:

$$\Delta F_m = F_{m-1} - F_m \approx \frac{F_m}{m} = \frac{(F_m)^2}{F_s}.$$

Относительная погрешность изменения частоты также увеличивается с увеличением частоты:

$$\frac{\Delta F_m}{F_m} \approx \frac{1}{m} = \frac{F_m}{Fs}.$$

Пример

Fm = 200 Гц, *Fs* = 8000 Гц. Тогда ΔF_m = 5 Гц.

Пример

Fo = 80 Гц, *Fs* = 8000 Гц получаем *Nmin* = 200 (размер окна 256).
Fo = 80 Гц, *Fs* = 11025 Гц получаем *Nmin* = 275 (размер окна 512).

Пример

Fs = 11025 Гц, *N* = 512 получаем *Fmin* = 42 Гц.

Можно сделать обратный расчет. Для наблюдения заданной частоты основного тона при фиксированной длине окна БПФ найти необходимую частоту дискретизации сигнала: *Fs = N_fft Fo*/2.

Пример

Fo = 80 Гц, *N* = 256, тогда *Fs* = 256 80 / 2 = 10240 Гц. Тогда можно выбрать стандартную частоту дискретизации сигнала 11025 Гц.

Между номерами коэффициентов и частотами существует соотношение $m = Fs/Fm$. Первому коэффициенту соответствует максимальная частота кепстра *Fmax = F1= Fs*, элементу с максимальным номером соответствует минимальная наблюдаемая частота.

Смысл частоты для кепстра заключается в следующем. Периодическому импульсному сигналу с частотой импульсов *Fm* и периодом *Tm = 1/Fm* соответствует дискретный периодический спектр с шагом пиков *Fm*.

Кепстр вычисляется с помощью преобразования Фурье, и коэффициентам кепстра соответствуют (как и у автокорреляционной функции) временные задержки $Tm = T \times m$ (T – временной шаг дискретизации сигнала). При этом задержка Tm в кепстре соответствует периодической структуре спектра с периодом Fm следования пиков. Это соотношение является основанием для представления кепстра в шкале частот: задержке *Tm* в кепстре можно сопоставить частоту $Fm = 1/Tm$.

3.3 Акустические явления в пространстве кепстров

Кепстр описывает структурные свойства спектра. Младшие коэффициенты кепстра описывают особенности формы спектра большей протяженности, старшие – мелкомасштабную структуру спектра.

Коэффициенты кепстра характеризуют форму логарифма спектра.

Разные коэффициенты описывают форму спектра различного масштаба в шкале частот:

$CC[0]$ –логарифм среднего значения спектра (мощность процесса);

$CC[1]$ – полупериод спектра, общий наклон спектра;

$CC[2]$ – полный период спектра (максимум – минимум – максимум);

$CC[m = Fs/1500 \ldots, CC[m = Fs/500]$ – широкие максимумы спектра (огибающая спектра);

$CC[m = Fs/400] \ldots CC[m = Fs/80]$ – временные события с периодами основного тона 80 .. 500 Гц;

$CC[m = Fn/\,Fo\_min] \ldots CC[m = N/2]$ – события в пространстве частот с малыми периодами, соответствующие мелким неравномерностям спектра, например, спектру шума;

$CC[m = N/2]$ – минимальный период вариации спектра (минимальный основной тон *Fo*).

Помимо формы спектра, коэффициенты кепстра характеризуют ещё и периодические акустические процессы с разным периодом следования. Акустические явления в пространстве кепстров представлены в таблице 3.1.

Таблица 3.1 — Акустические явления в пространстве кепстров

Частота *Fo* периода спектра, Гц	Период *To*, *мс*	Номера компонент, *m*	Акустические процессы, решаемые задачи
Fmax		0	Средняя энергия логарифма спектра
Fs		1	Общий наклон спектра
Fs/2		2	Максимальная частота основного тона
больше и равно 800	меньше 1,25	2..15	Крупномасштабная форма спектра речи, шума
400..800	меньше 2,5	10..25	Музыка, голоса птиц, сирена автомобиля
80..380	2..2,5	20..125	Основной тон речи, звука работающих механизмов
меньше 100	больше 10	больше 100	Эхо, резонанс, реверберация в небольшом помещении
50	20	*Fs*/25	Периодическая сетевая наводка
меньше 40		больше 200	Случайные мелкоразмерные флуктуации спектра
Fmin		*N*/2	Минимальная частота наблюдаемого основного тона

Нулевой коэффициент *CC*[0] соответствует вычислению с максимальным периодом или среднему значению логарифма спектра.

Коэффициенты *CC*[1], *CC*[2],… соответствуют постепенно уменьшающимся (от максимального) периодам амплитуды спектра (назовём частотами ОТ по аналогии со спектром речевого сигнала). По сути, это структурные коэффициенты формы, соответствующие периодам вариации амплитуды спектра (например, формантам спектра речевого сигнала). По мере уменьшения периода (возрастания частоты) этих вариаций они перестают соответствовать вариациям формант и начинают соответствовать вариациям спектра, связанным со спектральными максимумами основного тона (начиная от частоты ОТ 400 Гц в сторону более частых вариаций вплоть до 80 Гц). Коэффициентам с максимальными номерами коэффициентов кепстра соответствуют минимальные частоты ОТ (80 Гц).

3.4 Кратковременный кепстральный анализ

Кратковременный кепстральный анализ (на кадрах сигнала) определяется следующими соотношениями:

$$CCx(m,k) = IDFT\{\ln|X(n,k)|^2\} = IDFT\{\ln(DFT\{x[i]\})^2\},$$

$$|X(n,k)|^2 = \exp[DFT\{CCx(m,k)\}].$$

В ряде приложений к кратковременным оценкам кепстра применяют операции фильтрации – ФНЧ:

$$LPF\{CCx(m,k)\} = (1-\beta)LPF\{CCx(m,k-1)\} + \beta\, CCx(m,k)$$

или операцию ФВЧ:

$$HPF\{CCx(m,k)\} = (1-\beta)HPF\{CCx(m,k-1)\} + CCx(m,k) - CCx(m,k-1),$$

где β – коэффициент сглаживания.

Кепстрограмма является графическим представлением кратковременных кепстров на последовательности кадров сигнала:

$$CC[m,k] = CC(m, L \times k \times T),$$

где k – индекс кадров;

L – шаг кадров;

T – шаг дискретизации сигнала по времени.

Соотношение между периодическими сигналами средними спектрами и кепстрограммой представлено на рисунке 3.2.

На рисунке 3.2 показаны осциллограммы периодических импульсных сигналов с разной частотой следования импульсов (103 и 356 Гц), средние спектры и кепстрограммы этих сигналов.

Увеличение частоты следования импульсов (частота основного тона) приводит к увеличению интервала между пиками спектра и увеличению значения частоты ОТ на кепстрограмме.

Частота ОТ на кепстрограмме равна интервалу частоты между пиками спектра и частоте следования импульсов.

На кепстрограмме белым цветом обозначаются отрицательные и нулевые значения кепстров, тёмным цветом – положительные значения.

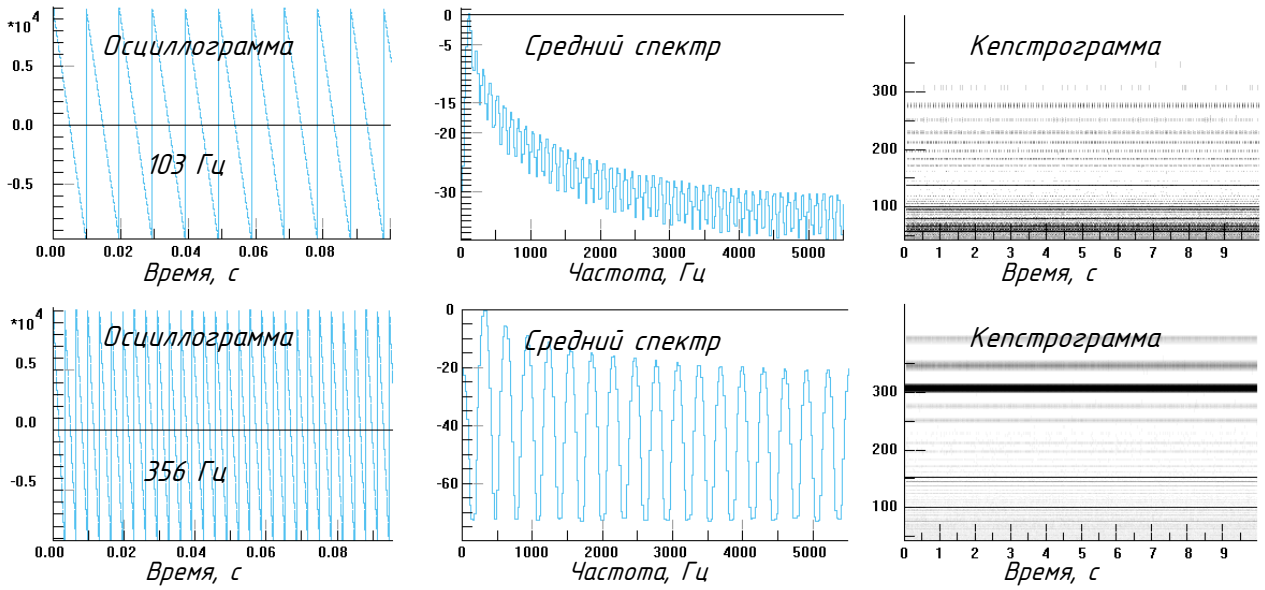


Рисунок 3.2 — Осциллограммы, средние спектры и кепстрограммы сигналов

Обычно кепстрограммы приводятся в координатах «номер коэффициента – время» или «частота-время». На рисунке 3.3 показана кепстрограмма речевого сигнала в координатах «время-частота».

На кепстрограмме заметна неравномерность шага шкалы частот (шаг растет с увеличением частоты). Также видна нижняя граница частот, определяемая длиной кадра анализа: *Fmin* = 2 ×11025/512 ≈ 43 Гц.

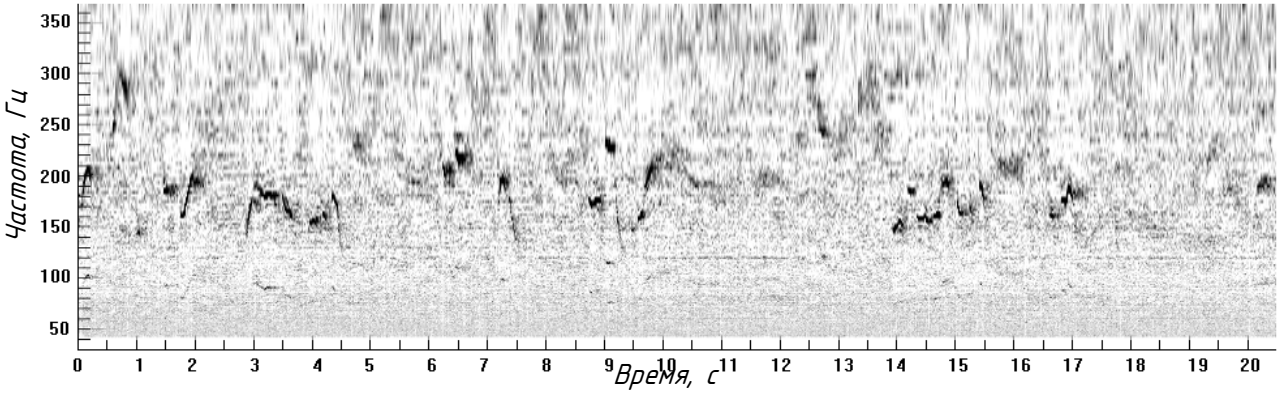


Рисунок 3.3 — Кепстрограмма речевого сигнала в координатах «время-частота» в области низких частот

На рисунке 3.4 показана кепстрограмма протяженного фрагмента речевого сигнала в координатах «время-частота». Тёмным цветом выделяется траектория основного тона речевого сигнала.

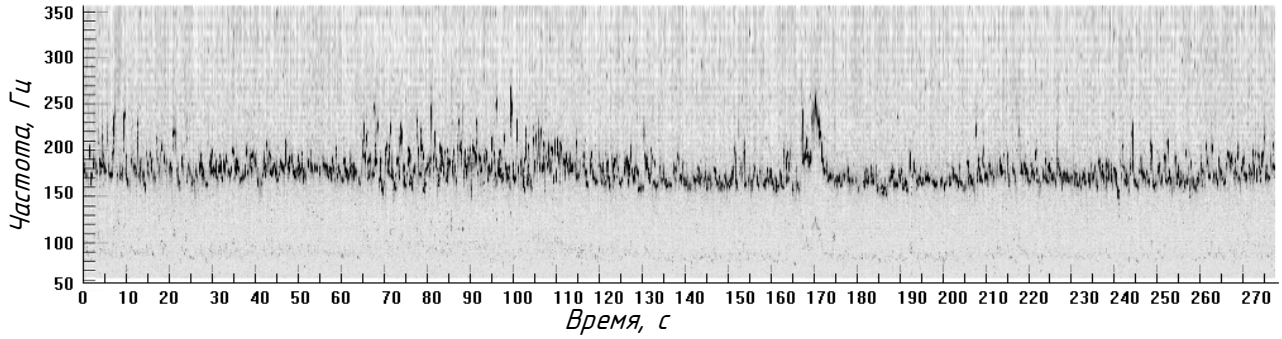


Рисунок 3.4 — Кепстрограмма речевого сигнала (*Fs* = 16 кГц, *N* = 512) в координатах «время-частота»

На рисунке 3.5 приведены средние кепстры периодических импульсных наводок с частотами следования импульсов 19, 22 и 50 Гц.

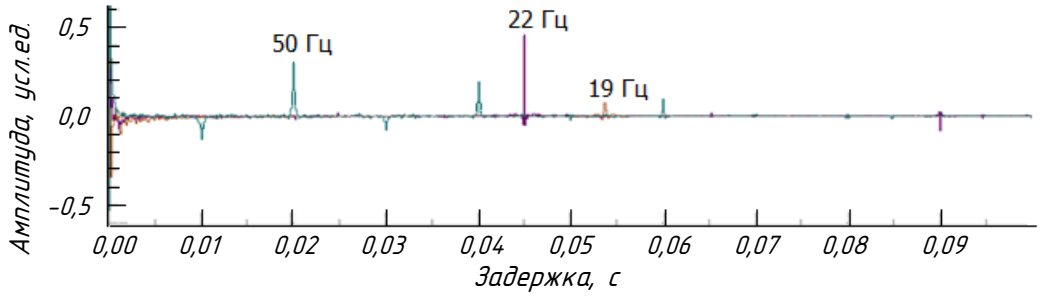


Рисунок 3.5 — Средние кепстры периодической импульсной наводки с частотами повторения 19, 22 и 50 Гц

Частоте 50 Гц соответствует пик кепстра с меньшей задержкой *To*, частоте 19 Гц соответствует пик кепстра с наибольшей задержкой. Максимумы кепстра повторяются на кратных задержках (×*2To*, ×*3To*,…), похожих на своеобразное «эхо». Это связано с формой импульсов сигнала и формой пиков его спектра, поскольку они не являются идеально гармоническими. На кепстограмме это «эхо» проявляется в виде траекторий с частотами кратными частоте основной траектории *F(t): F(t)/2, F(t)/3*,…

3.5 Применение кепстров

Кепстры хорошо описывают периодические формы спектров, поскольку основаны на разложении в периодические функции.

Также кепстры позволяют существенно сократить размерность представления спектров. Для описания крупномасштабной формы спектра требуется несколько младших коэффициентов кепстра.

Соотношение между спектрами и кепстрами заключается в следующем. Спектр эффективно представляет тональные временные процессы в частотной области. Кепстр описывает периодические (полигармонические) временные процессы.

Бóльшая часть акустических процессов (речь, музыка и др.) имеет полигармоническую структуру и представляется в спектре бóльшим числом равноотстоящих пиков. Анализировать и выполнять фильтрацию таких процессов в спектральном пространстве не всегда удобно.

В пространстве кепстров такие процессы имеют компактную структуру – как правило, в виде одного ярко выраженного пика. Поэтому обработка таких сигналов в пространстве кепстров может оказаться более удобной.

В пространстве кепстров можно решать различные задачи фильтрации и анализа аудиосигналов. Перечислим некоторые из этих задач:

– выравнивание формы спектра;
– устранение акустического эха;
– разделение дикторов по значению основного тона;
– подавление электрических наводок;
– подавление тональной музыки;
– выделение формант речевых сигналов;
– выделение основного тона речевых сигналов.

Обзор применений кепстров для обработки речевых сигналов приведен в [2, 3], для анализа работы механизмов — в [1].

Литература

1. Randall R. B. A History of Cepstrum Analysis and its Application to Mechanical Problems // International Conference at Institute of Technology of Chartres, France, October 29 – 30, 2013, pp 11 – 16.

2. Рабинер Л. Р., Шафер Р.. Цифровая обработка речевых сигналов // М.: Радио и связь, 1981. – 496 с. Гл. 7. Гомоморфная обработка речи

3. Оппенгейм А. В., Шафер Р. В. Цифровая обработка сигналов // М.: Связь, 1979.

Вопросы и задачи

1. Какому преобразованию сигнала соответствует вычитание (обнуление) нулевого коэффициента кепстра?

2. Какие коэффициенты кепстра (младшие или старшие?) описывают форманты речевого сигнала?

3. Какая операция в пространстве кепстров соответствует свертке сигналов во временной области?

4. Как изменится кепстр сигнала, если увеличить амплитуду сигнала в «е» (экспонента) раз?

5. Как минимальная наблюдаемая в кепстре частота связана с длиной окна ДПФ и частотой дискретизации сигнала?

6. Какому коэффициенту кепстра соответствует частота основного тона *Fo* сигнала, дискретизированного с частотой *Fs*?

7. Какой операции в пространстве кепстров соответствует временной сдвиг ΔT сигнала?

Глава 4
Характеристики и модели речевых сигналов

Материал, относящийся к данному предмету, чрезвычайно обширен. По теме речи и речеобразования написаны целые книги. Классической является книга Фланагана [1], хотя к настоящему времени она является несколько устаревшей, поскольку относится к «аналоговой эпохе».

Основной целью этой главы является введение двух новых понятий – *основного тона* и *формант*, а также методов их оценки. В остальном материал представляет лишь краткое введение в предмет.

4.1 Звуковой состав речевых сигналов

4.1.1 Общая характеристика речевых сигналов

Звуки речи образуются под действием давления воздуха, поступающего из лёгких, модуляции вибрирующих голосовых складок и резонансов голосового тракта, возникающих при выталкивании воздуха через губы и нос.

Речь является чрезвычайно информационно богатым сигналом, использующим частотную, амплитудную и временную модуляцию (движение резонансов гармоник и шума, интонацию основного тона, интенсивность и длительность звуков). Речь передает информацию о произносимых словах, личности диктора, акценте, эмоциях, стиле речи, состоянии здоровья диктора.

В основном вся эта информация может быть передана по телефонному каналу в диапазоне 4 кГц. Энергия речи выше этого диапазона в основном обеспечивает качество воспринимаемого звука.

Различные области и методы анализа РС направлены на извлечение этой информации.

Основные сферы применения анализа акустических характеристик РС следующие:

– автоматическое распознавание речи;
– идентификация дикторов;
– синтез речи;
– медицинские приложения;
– выделение речи в шумах;
– распознавание эмоций, состояния диктора;
– распознавание языка;
– кодирование речи в системах связи.

Методы извлечения этой информации основываются на оценке характеристик кратковременного анализа РС. В предыдущих главах были рассмотрены временные и спектральные методы анализа речевых и акустических сигналов. В данной главе основное внимание уделено специфическим характеристикам РС – основному тону и формантам, а также методам их оценки.

Изучением речи с позиций акустики занимаются две науки:

– *фонетика* – раздел лингвистики, изучающий звуки речи и звуковое строение языка – слоги, звукосочетания, закономерности соединения звуков в речевую цепочку;

– *речевая акустика* – раздел общей акустики, изучающий структуры речевого сигнала, процессы речеобразования и восприятия речи у человека.

Вопросы речеобразования подробно изложены в книгах [1–5].

4.1.2 Акустическое кодирование звуков речи

Звуки речи можно описать с разных позиций:

– восприятие;

– артикуляция;

– акустическая классификация.

С позиции *восприятия* классификация воспринимаемых звуков основывается на понятии фонемы. *Фонема* – это минимальная смыслоразличительная единица звукового строя данного языка, с помощью которой различаются и отождествляются слова и словосочетания (например, фонемы «*д*» и «*т*» в разнослышимых словах «*дом*» и «*том*»). Понятия фонема и звук речи не совпадают, так как фонема может состоять не только из одного звука [5].

Отметим также ещё два момента [4]:

1. фонема не есть физическая реализация звука, а является представлением звука в сознании (абстракцией);

2. фонема воплощает идею атомарности, примененную к субъективному представлению о речи.

С позиций *артикуляции* центральным моментом является анализ звуков с позиции их формирования органами речи.

Акустическую классификацию звуков разработали Р. Якобсон, Г. Фант и М. Халле [6]. Универсальная *акустическая классификация звуков* построена на учёте различных акустических параметров, включающих 12 бинарных признаков [3]. Акустическая классификация звуков позволяет увидеть многие особенности звуков, но она разработана менее детально и используется не так широко, как артикуляционная классификация.

Пример акустической классификации (кодирование фонем различными акустическими процессами):

– тональные – длительные порции сигнала с локализацией частот;

– переходные – широкополосный сигнал, локализованный во времени;

– шум с гладким спектром;

– переходные процессы с изменением спектрально-временных свойств;

– переходы гласная-согласная, согласная-гласная;

– гласные – положение формант и антиформант;

– согласные – джиттер, шиммер, квазигармонический компонент шума.

Для анализа речевых сигналов используется целый ряд характеристик. Вариант схемы анализа РС приведен на рисунке 4.1.

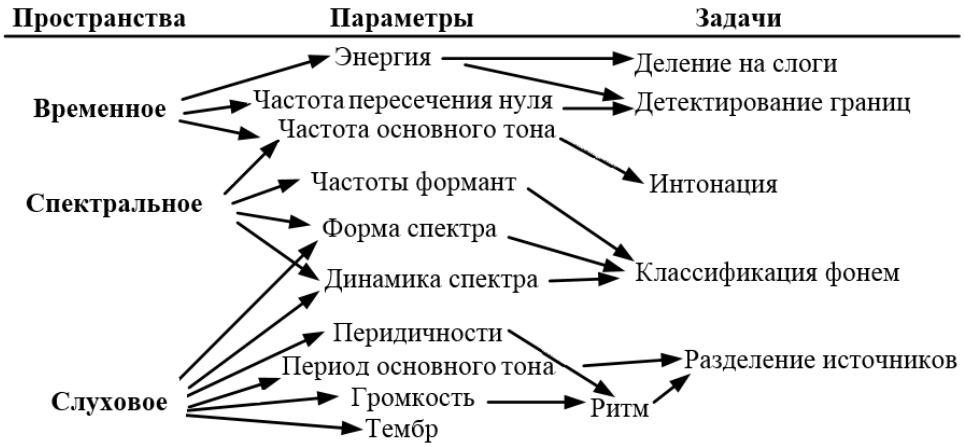


Рисунок 4.1 — Общая схема и задачи анализа речевых сигналов

4.1.3 Гласные и согласные

В первом приближении все звуки речи можно разделить на две большие группы: *гласные* (*vowels*) и *согласные* (*consonants*). Согласные звуки несут основную смысловую нагрузку, гласные – основную эмоциональную нагрузку. Согласные в свою очередь делятся на щелевые (глухие и звонкие) и взрывные (глухие и звонкие).

Пример акустического кодирования различных звуков речи в осциллограмме и спектрограмме приведен на рисунке 4.2.

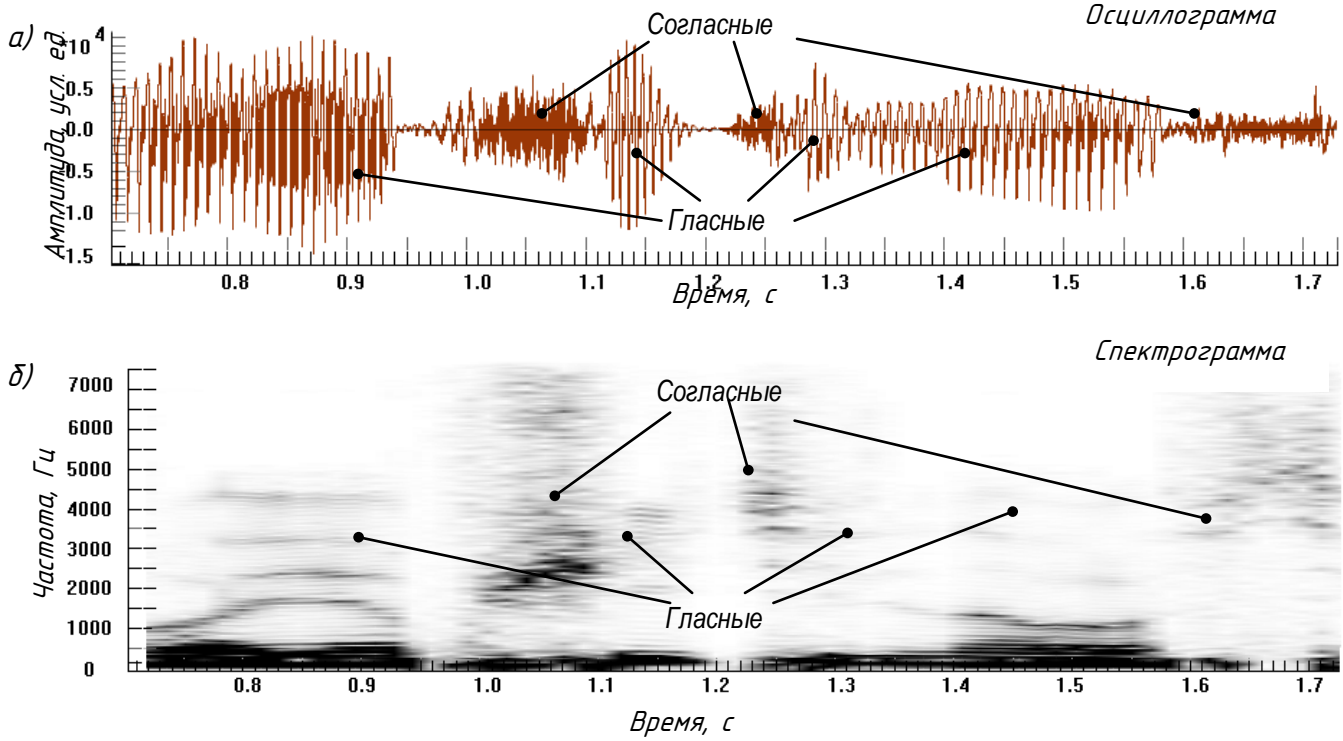


Рисунок 4.2 — Акустическое кодирование гласных и согласных звуков

Из рисунка видно, как кодируются звуки в частотно-временном измерении.

Гласные и согласные различаются способом формирования звука.

Гласные представлены квазистационарными порциями сигнала. Информация передается относительной для каждой частоты энергией (приблизительно постоянной на протяжении гласной). Гласные всегда являются тональными звуками.

Согласные передают информацию с помощью модуляции энергии, как по времени, так и по частоте. Бóльшая часть речевой информации передается с помощью временной модуляции энергий спектра на каждой частоте. Согласные могут быть тональными и не тональными.

4.1.4 Частотные свойства речевых сигналов

Разные звуки речи занимают различные области спектра (рисунок 4.3).

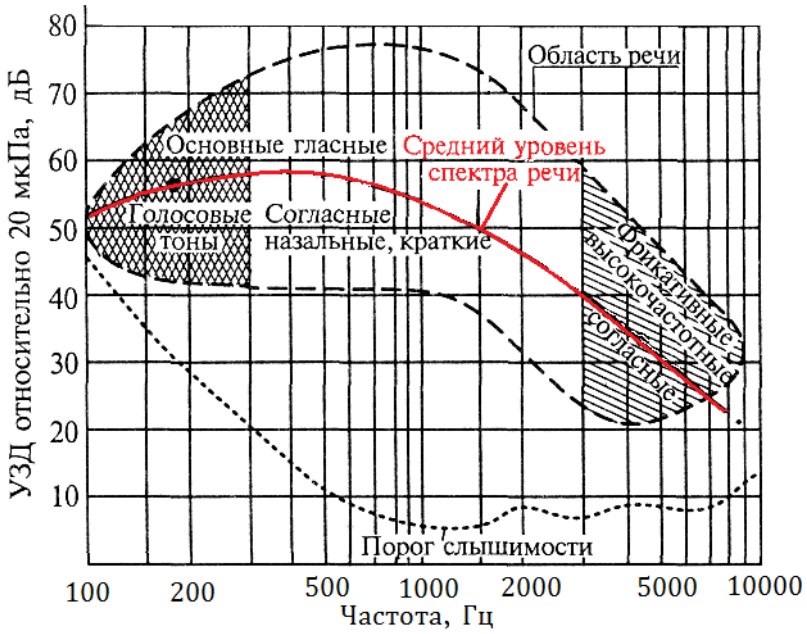


Рисунок 4.3 — Спектральные области звуков речи [7]

Спектр отдельных звуков может выходить за границы диапазона слуха человека. На рисунке 4.4 представлена спектрограмма фрагмента РС с частотой дискретизации 48 кГц.

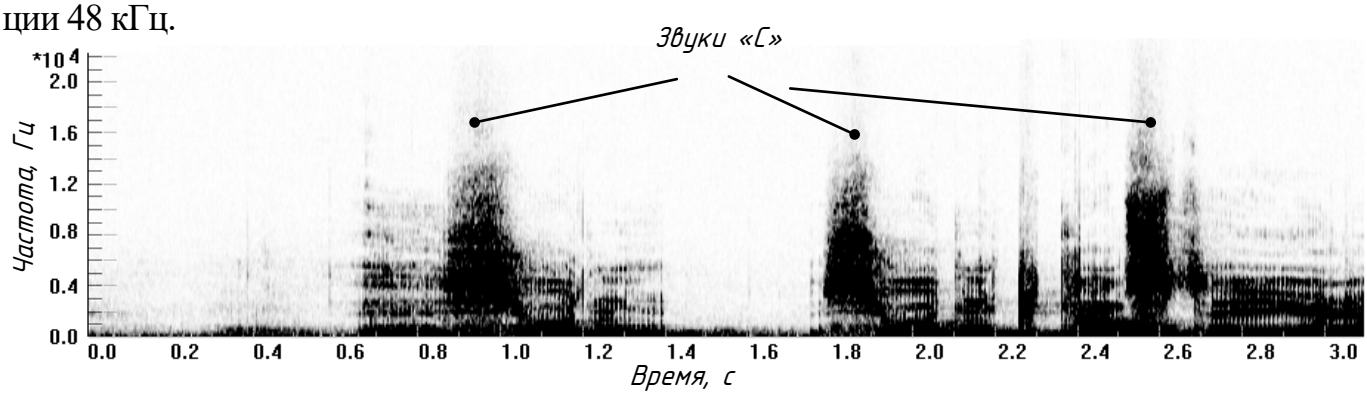


Рисунок 4.4 — Спектрограмма (окно 256 точек) речевого сигнала

Из спектрограммы видно, что основная энергия тональных звуков сосредоточена в диапазоне 5 кГц, в то время как энергия звуков «С» находится в диапазоне до 24 кГц.

4.1.5 Временные свойства речевых сигналов

В работе [8] рассмотрены методы измерения следующих временных характеристик речи:

– длительность звука речи;
– длительность пауз речи между словами;
– длительность пауз речи между фразами;

– скорость звуков речи, звуков/с;

– темп речи, слов/мин;

– плотность речи, Р % – отношение времени наличия звука к полному времени речевого сигнала;

– скорость изменения уровня громкости основного тона, дБ/с;

– скорость изменения частоты основного тона, Гц/с.

Числовые значения некоторых временных характеристик речевого сигнала приведены в таблице 4.1.

Таблица 4.1 — Временные свойства речевого сигнала

Длительность	Речевые события
5 мс	Длительность взрывных согласных (П, Б,…)
2–12,5 мс	Длительность периодов основного тона
10–25 мс	Длительность невокализованных звуков. Изменение характеристик речевого тракта незначительное
25–40мс	Гласные звуки: временной интервал, на котором речь может рассматриваться как стационарный процесс
40–100 мс	Длительность звуков вокализованной речи
25–250 мс	Интервал модуляции интенсивности звуков
0,1–0,2 с	Отдельные слова
1–2 с	Отдельные фразы
5–7 с	Сентенции, последовательности фраз, ритм дыхания
50 с	Последовательность сентенций (тема фрагмента разговора)
5 мин	Тема в разговоре
25 мин	Разговор

4.2 Основной тон и его оценка

Основной тон (ОТ, *fundamental frequency*) – кратковременная частота колебаний голосовых складок (связок) диктора.

4.2.1 Период и частота основного тона

Период (*pitch*) и частота ОТ связаны следующими соотношениями:

$$F_0[\text{Гц}] = \frac{1}{T_0[c]}\ , T_0[c] = \frac{1}{F_0[\text{Гц}]}.$$

Пример

Fo = 100 Гц → *To* = 10 мс

4.2.2 Представление основного тона в осциллограмме и спектре

На рисунке 4.5 представлена осциллограмма фрагмента тонального речевого сигнала и амплитудный спектр, вычисленный по данному фрагменту.

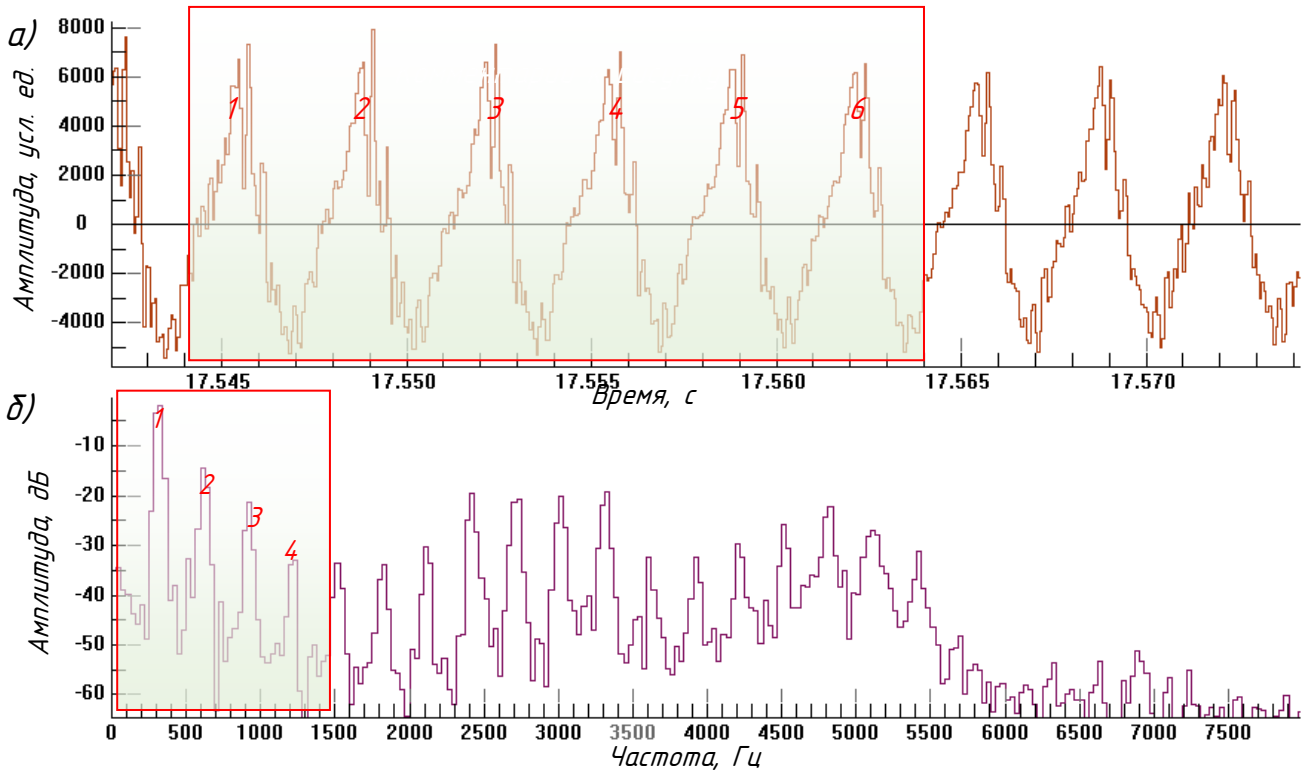


Рисунок 4.5 — Представление основного тона: а) на осциллограмме, б) мгновенный амплитудный спектр фрагмента речевого сигнала

По осциллограмме и спектру можно вычислить период и частоту основного тона. В осциллограмме на временной интервал 17,545–17,565 с приходится 6 периодов. Следовательно, период ОТ To = 20 мс/6 ≈ 3,33 мс. В спектре на интервал частот 0–1200 Гц приходится 4 пика. Таким образом, частота ОТ Fo = 300 Гц. Оценки, полученные из спектра и осциллограммы совпадают: $Fo = 1/To$.

4.2.3 Кодирование информации основным тоном

Основной тон является носителем разнообразной информации о дикторе:
– гендерная принадлежность (мужчина, женщина, ребенок);
– интонация (вопросительная, утвердительная, …);
– возраст;
– акцент (вологодское «о» и пр.);
– индивидуальные признаки говорящего;
– эмоции (стресс, смех, удивление и др.);
– физиологическое и психологическое состояние;
– состояние здоровья.

Информация кодируется временной динамикой основного тона, длительностями и энергией тональных участков.

Далее приведены несколько примеров кодирования информации основным тоном.

На рисунке 4.6 приведен пример спектрограммы и кепстрограммы фрагмента диалога женщины (3,7–5,4 с) и мужчины (6,4–8,0 с).

Из кепстрограммы видно, что траектория основного тона женщины находится в интервале 190–390 Гц, мужчины – в интервале 100–180 Гц.

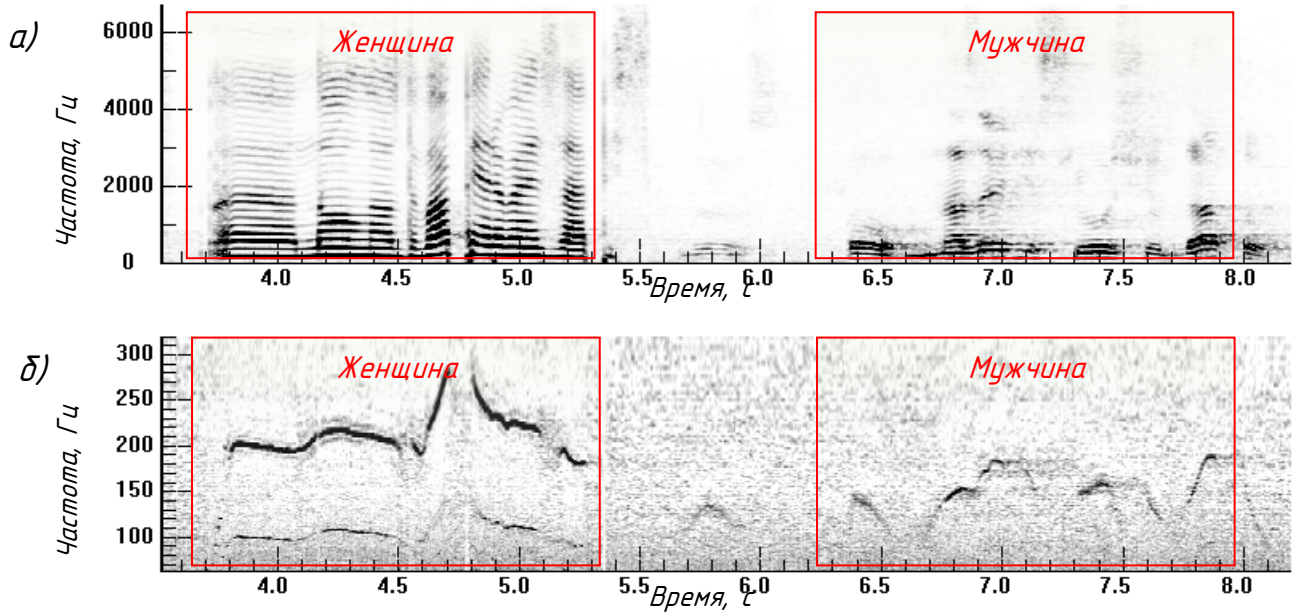


Рисунок 4.6 — Представление траектории ОТ а) на спектрограмме
б) на кепстрограмме

Средние значения и интервалы частот основного тона мужчин, женщин и детей приведены в таблице 4.2.

Таблица 4.2 — Средние значения и средние интервалы частот ОТ

Гендерная принадлежность	Средние значения, Гц		
	Среднее значение	Минимальный интервал	Максимальный интервал
Мужчины	125	80	200
Женщины	225	125	350
Дети	300	200	500

В отдельных случаях, значения частоты основного тона могут существенно выходить за интервалы частот, приведенных в таблице.

На рисунке 4.7 показана спектрограмма фрагмента речевого сигнала в телефонном канале. На временном интервале 4,6–4,8 с частота основного тона быстро возрастает до значений более 600 Гц.

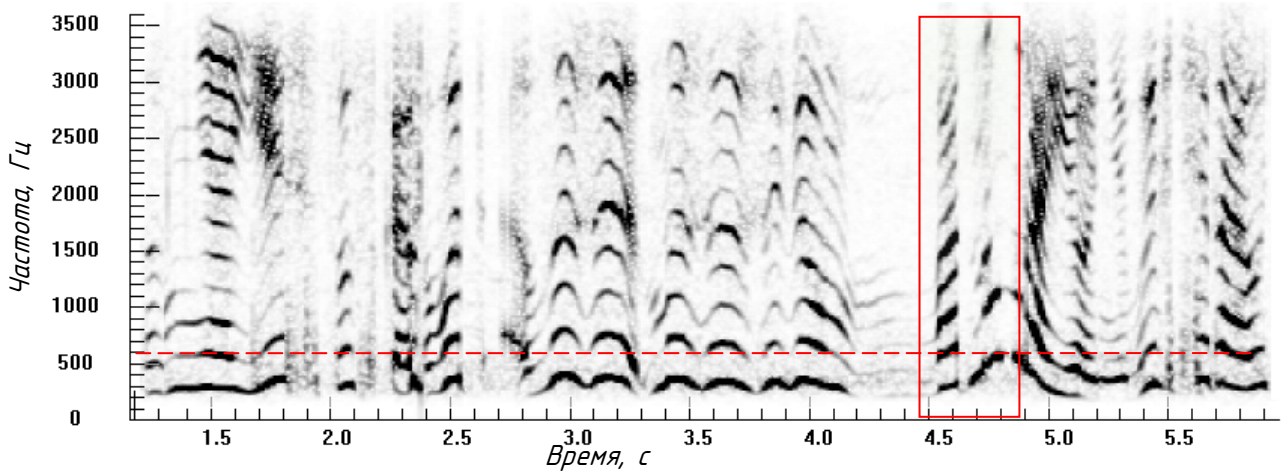


Рисунок 4.7 — Спектрограмма фрагмента речевого сигнала в телефонном канале

На рисунке 4.8 приведен пример траектории основного тона на кепстограмме в диалоге: вопрос 1,0–1,7 с, ответ («да») 2,4–2,7 с.

Вопросительная интонация диктора кодируется повышением частоты ОТ.

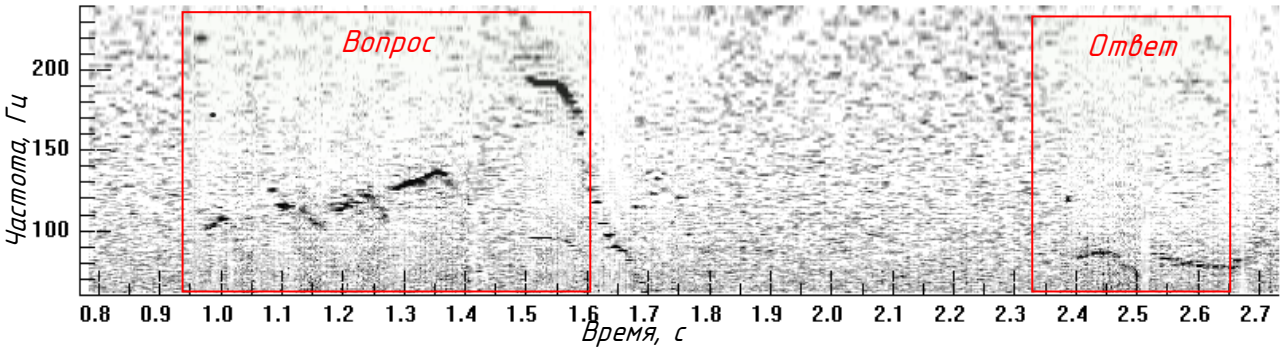


Рисунок 4.8 — Спектрограмма тонального фрагмента речевого сигнала с вопросительной интонацией

Пример кодирования информации (смех) основным тоном приведен на рисунке 4.9.

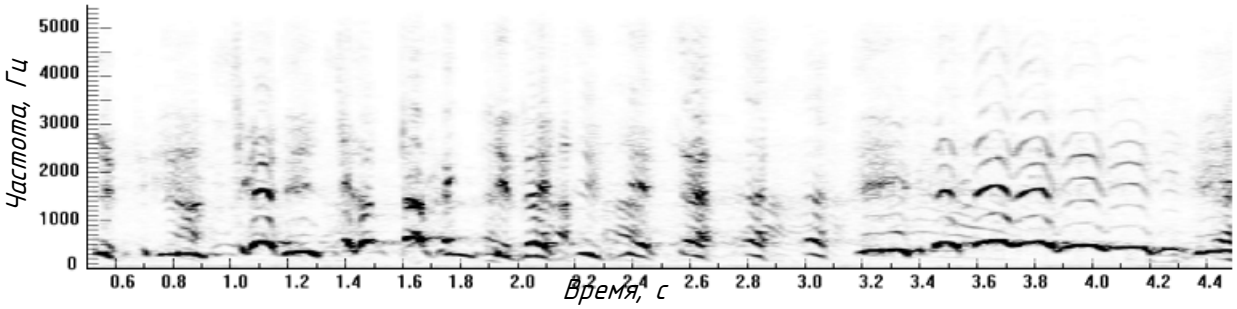


Рисунок 4.9 — Спектрограмма смеха мужчины

Из спектрограммы видно, что траектории основного тона смеющегося человека существенно отличаются от траекторий обычной речи.

Из рассмотренных примеров ясно, что основной тон является информативной характеристикой при анализе различных акустических сигналов.

4.2.4 Методы и алгоритмы оценки частоты основного тона

Классические методы оценки частоты ОТ рассмотрены в книгах:

– с использованием функции автокорреляции [2];

– по временным периодам речевого сигнала [9];

– с использованием кепстра [10];

– обзор методов оценки ОТ приведен в [11].

При построении траектории ОТ могут возникать определенные трудности: траектория ОТ может прерываться (не наполняется энергией либо заполняется другими звуками). Для построения непрерывной траектории ОТ (мелодический контур) РС применяется интерполяция с помощью медианного и сглаживающего фильтров [2].

Например, алгоритм детектирования ОТ с использованием кепстра состоит из следующих шагов:

– вычисление кепстра каждые 20–30 мс;

– поиск пика среди коэффициентов, соответствующих интервалу ОТ;

– если амплитуда пика больше порога, то речь тональная и вычисляется частота ОТ;

– если нет, то речь не тональная.

4.3 Форманты и их оценка

4.3.1 Форманты

Форманты — это максимумы частотной огибающей спектра.

Форманты речевого сигнала характеризуются частотой и амплитудой максимумов спектра, их шириной и временными траекториям.

Форманты соответствуют резонансам голосового тракта.

Форманты кодируют как лингвистическую информацию, так и информацию о дикторе.

На рисунке 4.10 показаны мгновенные спектры на кадрах тонального и не тонального речевого сигнала (окно 512 отсчетов), и их частотные огибающие, вычисленные с окном 128 отсчетов.

Частоты формант тонального РС, рис. 4.10 а):

F1= 500 Гц, F2= 2700 Гц, F3 = 4700 Гц, F3 = 6500 Гц.

Частоты формант нетонального РС, рис. 4.10 б):

F1= 1000 Гц, F2= 2300 Гц, F3 = 3500 Гц, F3 = 5000 Гц.

Отметим, что изменение амплитуды между формантными максимумами и межформантными минимумами составляют примерно 12 дБ, что соответствует перепаду огибающей амплитудного спектра в 4 раза, спектра мощности – в 16 раз.

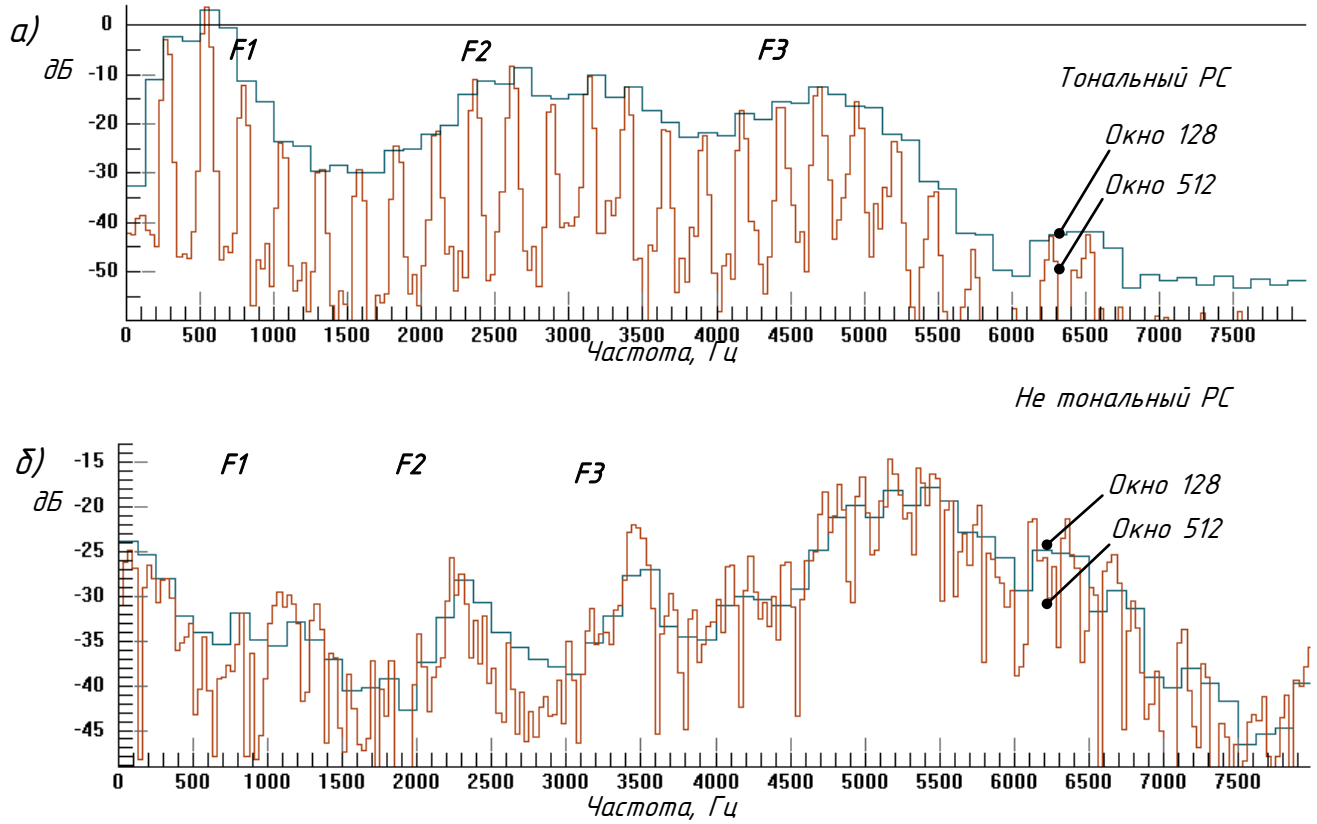


Рисунок 4.10 — Мгновенные амплитудные спектры тонального и не тонального РС с окнами 512 и 128 отсчетов

4.3.2 Форманты гласных: кодирование информации

Гласные звуки кодируются частотами двух младших формант F1 и F2. Соотношение F1 и F2 по Фанту следующее (таблица 4.3).

Таблица 4.3 — Положение формант F1 и F2 гласных звуков

Форманта, Гц	Звуки					
	[и]	[э]	[ы]	[а]	[о]	[у]
F1	230	420	285	630	500	240
F2	2220	1960	1480	1070	860	610

По Л. В. Бондарко [12] это соотношение имеет следующий вид (рисунок 4.11).

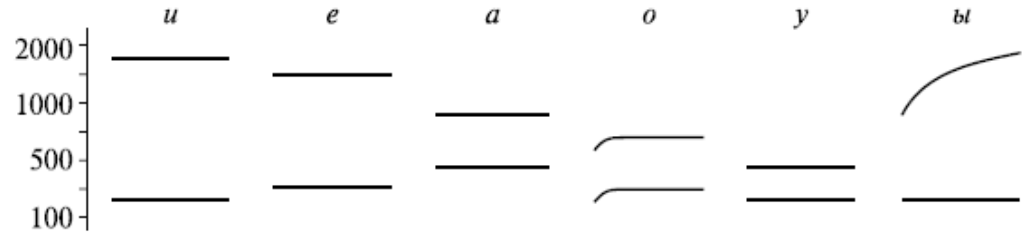


Рисунок 4.11 — Положение формант F1–F2 гласных звуков

Аналогичный принцип двухчастотного звукового кодирования символов (*dual-tone multi-frequency, DTMF*) применяется для автоматической телефонной сигнализации между устройствами. На рисунке 4.12 приведен пример спектрограммы сигналов *DTMF*.

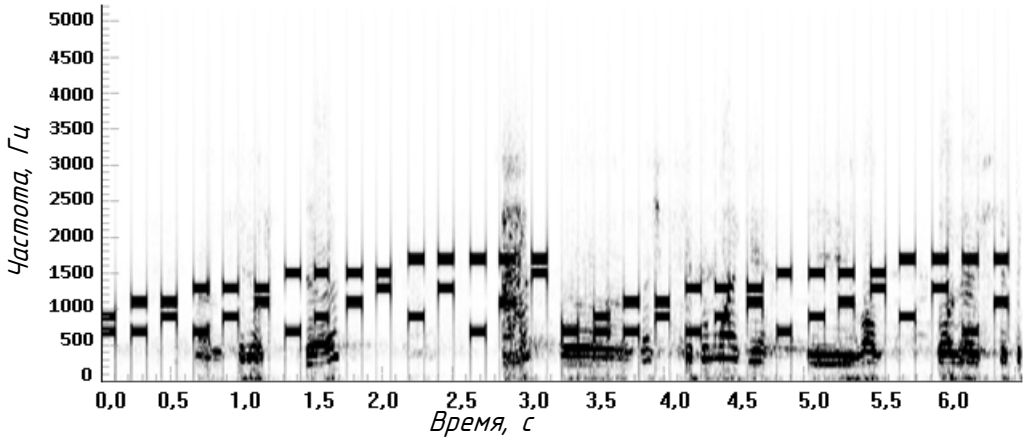


Рисунок 4.12 — Спектрограмма сигнала *DTMF*

4.3.3 Методы и алгоритмы оценки формант и формантных траекторий

Основными методами оценки формант и формантных траекторий являются:

– сглаживание линейного спектра Фурье (широкополосная спектрограмма Фурье);

– частотный отклик коэффициентов линейного предсказания [2];

– кепстральная фильтрация [10].

На рисунке 4.13 приведены широкополосная спектрограмм Фурье и спектрограмма, полученная на основе модели линейного предсказания речевого сигнала с частотой дискретизации 11025 Гц.

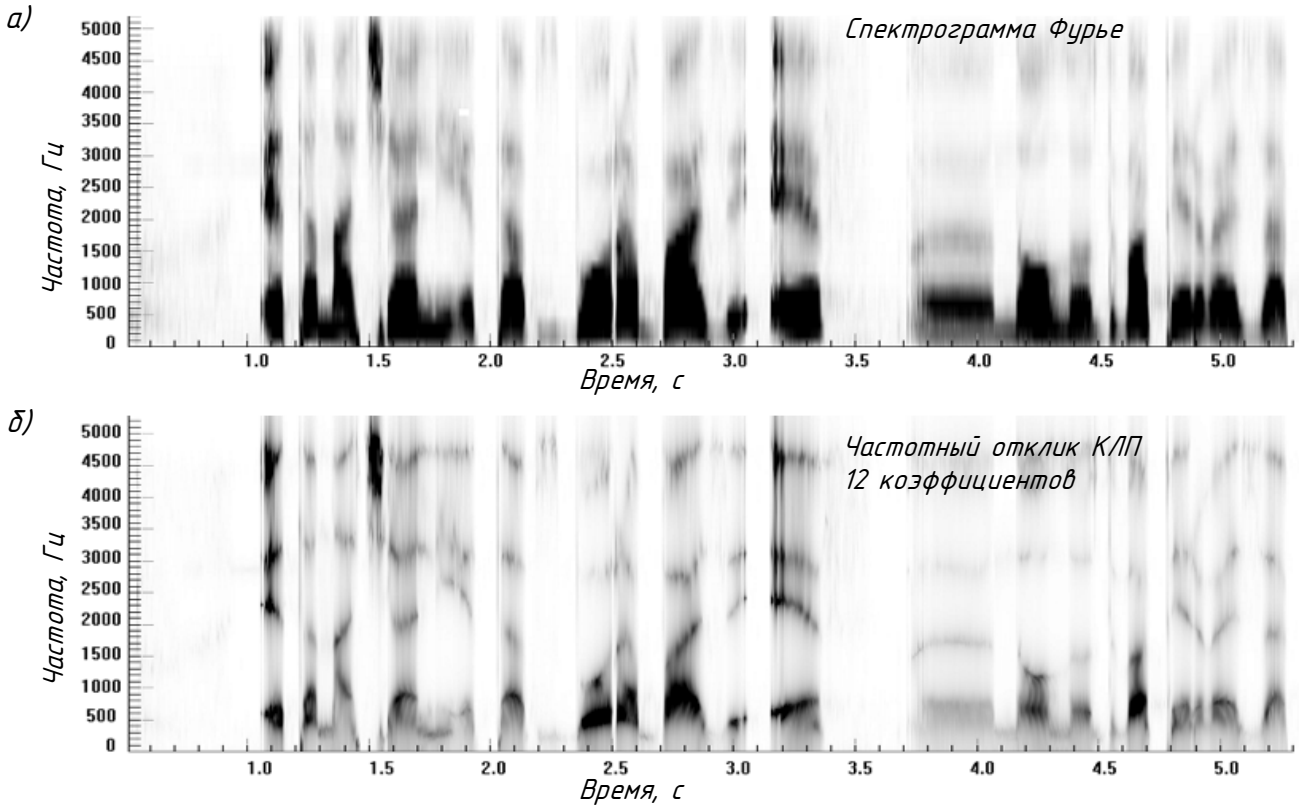
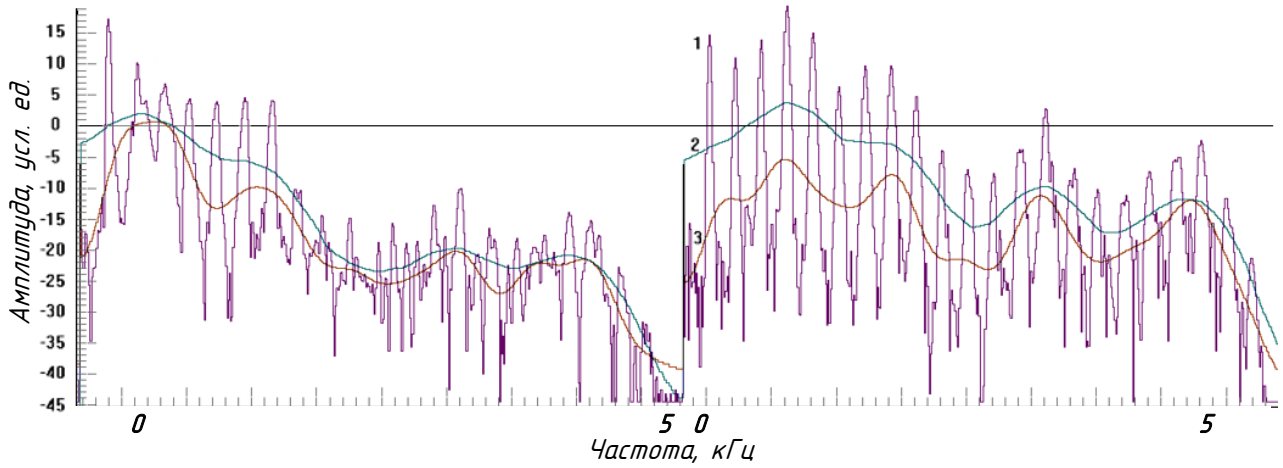


Рисунок 4.13 — Спектрограммы фрагмента сигнала с окном 64 точки

Кепстральная фильтрация

Идея кепстральной фильтрации заключается в преобразовании спектра в кепстр, обнулении коэффициентов кепстра с частотами выше граничной (максимальной) частоты ОТ, обратном преобразовании кепстра в спектр и оценке формантных максимумов в модифицированном спектре).

На рисунке 4.14 показаны спектры на двух соседних кадрах речевого сигнала: исходного сигнала, сглаженные спектры (с окном 500 Гц), спектры с обнулением коэффициентов кепстра в диапазоне частот от 120 Гц до 500 Гц).



1 – спектр исходного сигнала, 2 – спектр, сглаженный окном 500 Гц,
3 – спектр после обнуления коэффициентов кепстра выше 500 Гц

Рисунок 4.14 — Спектры речевого сигнала на двух последовательных кадрах

Из приведенного примера следует, что формантные максимумы, вычисленные разными методами, имеют близкие значения, но не совпадают.

Формантные траектории

Формантные траектории – последовательность значений частот формант: F1(*t*), F2(*t*), F3(*t*). Задача заключается в преобразовании частот максимумов спектра на кадрах в последовательности значений частот формант с соответствующими номерами.

Основные проблемы построения траекторий формант связаны со следующими моментами:

– необходимо исключить случайные максимумы спектров на кадрах;

– необходимо устранить случайные прерывания траекторий (пропадание максимумов спектров на кадрах);

– необходимо правильно пронумеровать форманты, например, в случае отсутствия максимума спектра, соответствующего форманте F3 необходимо присвоить значения частот формантам с номерами F1, F2, F4.

На рисунке 4.15 показана последовательность исходных максимумов спектров на кадрах.

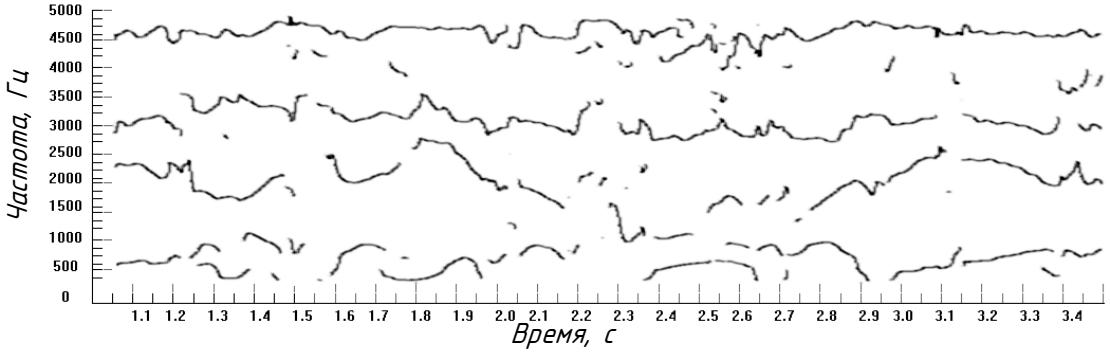


Рисунок 4.15 — Максимумы спектров фрагмента речевого сигнала

На рисунке 4.16 показаны формантные траектории фрагмента речевого сигнала, вычисленные с использованием метода сглаживания линейных спектров (рис. 4.16 а)) и с использованием модели линейного предсказания (рис. 4.16 б)).

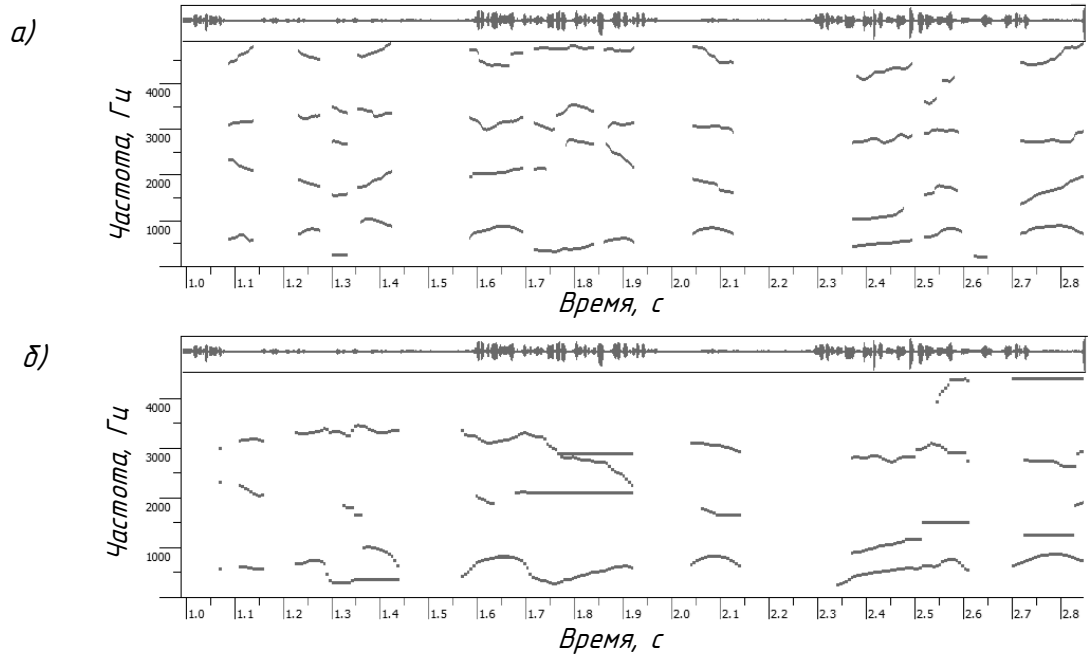


Рисунок 4.16 — Формантные траектории фрагмента речевого сигнала

Из приведенных примеров следует, что формантные максимумы, вычисленные разными методами, не всегда совпадают. Кроме того, некоторые формантные максимумы пропадают, а другие являются ложным. Вследствие этого задача оценки формантных траекторий не является тривиальной.

4.4 Модели речевых сигналов

4.4.1 Задачи моделирования речевого сигнала

Моделирование речевого сигнала применяется для решения разных задач:

– представление РС в виде компактного набора параметров для задач кодирования, распознавания, синтеза и компенсации искажений;

– выделение параметров, которые описывают разные характеристики РС, такие как акцент, эмоции, возраст и др.;

– детектирование речевых событий (гласные, согласные, переходы гласная-согласная, согласная-гласная) и оценка параметров.

Для кодирования наиболее популярными коммерческими системами являются системы на основе моделей линейного предсказания.

Для систем синтеза речи часто используется модель «гармоника + шум».

Для систем распознавания речи и распознавания дикторов наиболее популярными признаками, выделяемыми из огибающих спектра, являются различные кепстральные признаки (*LFCC, MFCC, LPCC*) и их временная динамика (производные 1-го и 2-го порядка).

Для задач распознавания эмоций важными признаками являются динамика ОТ и энергии.

4.4.2 Задача анализа и синтеза речевого сигнала

Целью *моделирования сигнала речевого источника* заключается в установлении соответствия между различными звуками и их акустическими признаками, то есть в параметрическом представлении существенных характеристик речевого сигнала. С практической точки зрения целесообразно использовать модели с минимальным числом параметров [2]. Моделирование предполагает решение двух задач.

Задача анализа РС: оценка параметров РС.

Задача синтеза РС: восстановление речевого сигнала по значениям параметров с использованием модели речевой системы.

Рассмотрим в качестве примера две модели.

4.4.3 Линейная модель речевого источника

Модель состоит из генератора сигнала возбуждения (тонального или шумоподобного), фильтра вокального тракта и фильтра губ.

Вокальный тракт представляется системой с медленно меняющимися параметрами. Предполагая, что параметры меняются медленно (приблизительно постоянны на отдельных участках), их можно вычислять на отдельных кадрах. Модель РС – сигнал возбуждения, пропущенный через цифровой фильтр с медленно меняющимися параметрами. При этом РС представлен параметрами модели вместо оцифрованного РС.

Схема линейной модели и последовательность формирования спектра тонального речевого сигнала приведены на рисунке 4.17.

Возбуждающий сигнал формируется генератором возбуждения (*excitation generator*) и далее модулирующий фильтр (*linear system*). Произнесение является динамическим процессом с изменяющимися параметрами генератора возбуждения и параметрами речевого тракта.

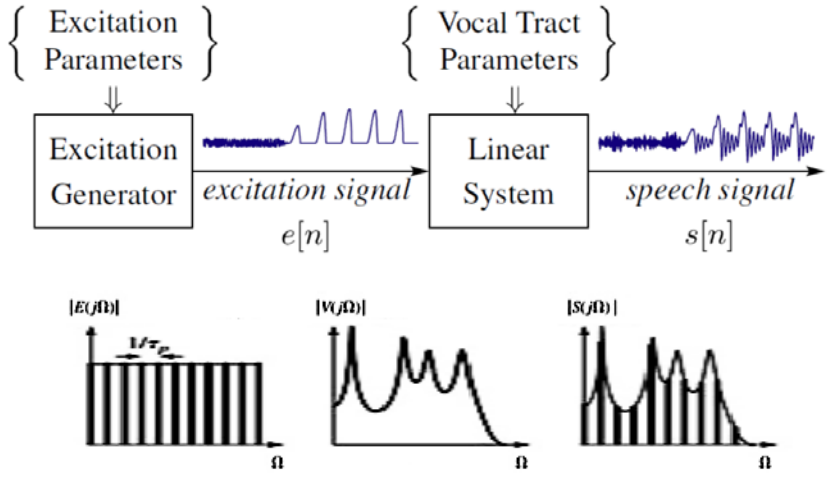


Рисунок 4.17 — Схема линейной модели речевого источника а), формирование спектра тонального речевого сигнала, спектры сигналов линейной модели б) [13]

Во временной области РС представляется в виде свертки *возбуждающего сигнала* $e[n]$, генерируемого связками, и *модулирующего фильтра* $h[n]$, являющегося характеристической функцией формы ротовой полости и артикуляторной характеристики:

$$s[n] = e[n] * h[n].$$

В частотно-временной области амплитудно-частотный спектр результирующих звуковых колебаний является произведением амплитудно-частотного спектра колебаний, создаваемых источником звука, и передаточной функции резонаторной системы речевого тракта:

$$S(f,t) = E(f,t)\,H(f,t),$$

где $H(f,t)$ – модулирующая передаточная функция голосового тракта;

$E(f,t)$ – спектр возбуждения речи.

Модулирующая передаточная функция представляет собой спектральную огибающую (форманты) речевого сигнала.

4.4.4 Модель линейного предсказания

Модель сигнала

Процесс речеобразования рассматривается как генерация возбуждающего сигнала (периодического или шумоподобного) и воздействие на него моделирующего фильтра линейного предсказания, управляющего формой и положением формант [2]:

$$s[n] = \sum_{k=1}^{M} s\,[n-k]\,a[k] + e[n],$$

где $a[k]$ – коэффициенты линейного предсказания;

$e[n]$ – возбуждающий сигнал.

Порядок модели при формантном анализе выбирают несколько больше частоты сигнала в килогерцах [2]:

$$M = Fs\,[\text{кГц}] + 4(5).$$

Анализ сигнала

Анализ исходного сигнала $s[n]$ заключается в оценке КЛП и остаточного сигнала $r[n]$, представляющего ошибку предсказания:

$$r[n] = s[n] - s[n] * a[n].$$

В z-представлении:

$$R(z) = S(z)\,A(z),$$

где функция (оператор) $A(z) = 1 - \sum_{k=1,\,M} a[k]\,z^{-k}$.

Минимизация ошибки линейного предсказания позволяет вычислить оценку КЛП на основе системы уравнений Юла–Уокера [2].

Синтез сигнала

Синтез (восстановление) моделируемого сигнала осуществляется через остаточный сигнал и оценку КЛП:

$$S(z) = \frac{R(z)}{A(z)} = R(z)H(z).$$

Передаточная функция $H(z)$ синтезирующего фильтра обратна частотной характеристике голосового тракта $A(z)$:

$$H(z) = 1/A(z).$$

4.4.5 Гармоническая модель речевого сигнала

Модель сигнала

Речевой сигнал представляется в виде суммы гармонического (тонального) и шумоподобного (нетонального) сигнала [14]:

$$x[n] = s[n] + u[n],$$

$$s[n] = \sum_{k=1}^{K} a[k]\cos(k\omega_0 nT) + b[k]\sin(k\omega_0 nT),$$

где $s[n]$ – тональный сигнал(гармоники);

$u[n]$ – шумоподобный (нетональный) сигнал;

ω_0 – фундаментальная частота ОТ;

$a[k]$, $b[k]$ – амплитуды гармоник.

Считаем, что число коэффициентов $a[k]$, $b[k]$ невелико и соответствует числу наблюдаемых гармоник.

Анализ сигнала

Считается, что величина ОТ определена. Тогда матрица косинусов и синусов ($n = 1, 2,..N$) известна: **A** – $N \times 2K$ матрица **A** = [**Acos Asin**].

Уравнение наблюдения:

$$\mathbf{X} = \mathbf{A}\,\mathbf{c} + \mathbf{U},$$

где **X** – вектор наблюдений (N отсчетов);

$\mathbf{c}^{\mathrm{T}}$ = [$a1\ a2\ldots aK\ b1\ b2..bK$] – вектор коэффициентов.

Вектор амплитуд гармоник может быть вычислен на основе метода наименьших квадратов. Определим вектор невязок между сигналом и моделью гармоник:

$$\mathbf{E} = \mathbf{X} - \mathbf{A}\,\mathbf{c}.$$

Функция квадратов ошибок определяется соотношением

$$\mathbf{E}\,\mathbf{E}^{\mathbf{T}} = (\mathbf{X} - \mathbf{A}\,\mathbf{c})\,(\mathbf{X} - \mathbf{A}\,\mathbf{c})^{\mathbf{T}}.$$

Минимизируя функцию квадратов ошибок по отношению к вектору **c**, получим

$$\hat{c} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\,\mathbf{X}\,.$$

Синтез сигнала

Зная значения коэффициентов, восстанавливаем по ним сигнал.

Оценка вектора отсчетов гармонической части сигнала: $\hat{\mathbf{Y}} = \mathbf{A}\,\hat{\mathbf{c}}$.

Оценка вектора отсчетов шумовой компоненты сигнала: $\hat{\mathbf{U}} = \mathbf{X} - \hat{\mathbf{Y}}$.

В заключение отметим, что моделирование речевых сигналов может выполняться в пространствах различных параметров.

Схема взаимосвязи между их основными представлениями приведена на рисунке 4.18.

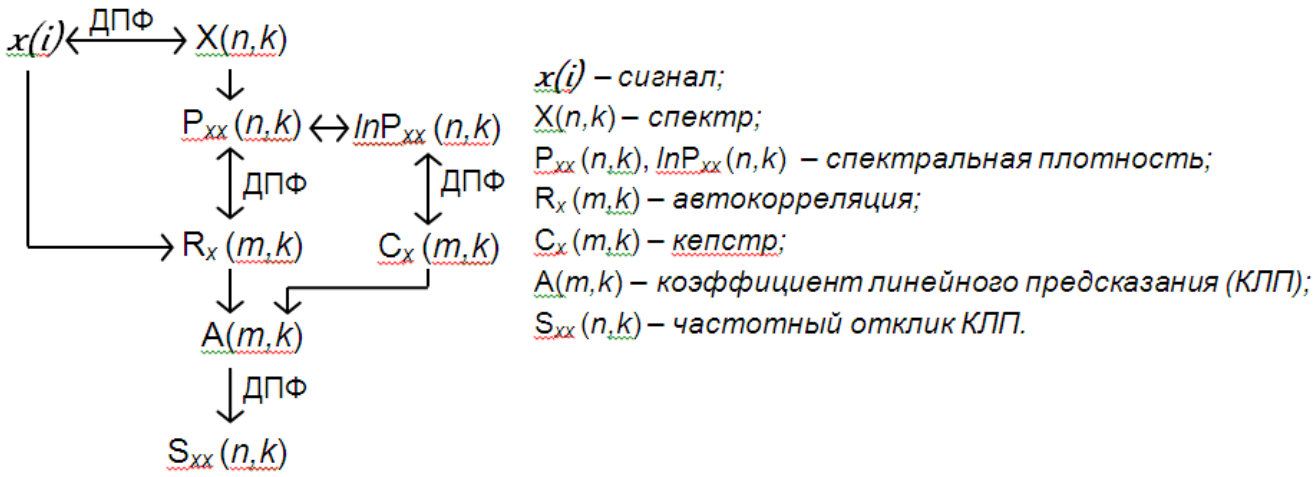


Рисунок 4.18 — Схема взаимосвязи между основными представлениями речевых сигналов

Литература

1. Фланаган Дж. Анализ, синтез и восприятие речи. — М.: Связь, 1968.

2. Маркел Дж. Д., Грей А.Х. Линейное предсказание речи. — М.: Связь, 1980. — 308 с.

3. Фант Г. Акустическая теория речеобразования. — М.: Наука, 1964. — 283 с.

4. Тампель И. Б., Карпов А. А. Автоматическое распознавание речи. Учебное пособие. – СПб.: Университет ИТМО, 2016. — 138 с.

5. Гатчин Ю. А. и др. Теоретические основы защиты информации от утечки по акустическим каналам: учебное пособие. // Новосибирск: СГГА, 2008. — 194 с.

6. Jakobson R., Fant G., Halle M. Preliminaries to speech analysis: The Distinctive Features and their correlates. – the MIT Press, 1952. – 64 p.

7. Руленкова Л. И., Смирнова О. И. Аудиология и слухопротезирование: Учеб. Пособие. // М.: Издательский центр «Академия», 2003. — 208 с.

8. Иванушкин С. А. и др. Основные спектральные характеристики речевых сигналов и их определение // Компьютерное моделирование в фундаментальных и прикладных исследованиях, 2021.

9. Рабинер Л., Гоулд Б. Теория и применение цифровой обработки сигналов. — М.: Мир, 1978. — 848 с.

10. Рабинер Л., Шафер Р. Цифровая обработка речевых сигналов // М.: Радио и связь, 1981. — 496 с.

11. Имамвердиев Я. Н., Сухостат Л. В. Подходы для оценки периода основного тона речевого сигнала в зашумлённой среде // Речевые технологии 1-2/2014, С.84-102.

12. Бондарко Л.В. Звуковой строй современного русского языка. – М.: Просвещение, 1977. — 175 с.

13. Rabiner L. R. and Schafer R. W. Introduction to Digital Speech Processing // Foundations and Trends in Signal Processing. Vol. 1, No.1–2 (2007) 194 P.

14. Павловец А. Н, Зубрыцки П., Петровский А. А. Гармоническая модель речевого сигнала: определение параметров и их квантование // Доклады БГУИР №4 (20), 2007, с. 20-34.

Вопросы и задачи

1. Каков частотный диапазон речевых сигналов? Каков диапазон основных голосовых тонов?

2. Гласные и согласные звуки. Чем они отличаются с точки зрения акустики? В каких частотных диапазонах сосредоточена энергия гласных и согласных звуков?

3. Основной тон и способы его оценки. Какую информацию содержит ОТ?

4. Форманты и способы их оценки. Какую информацию содержат форманты?

5. Модель линейного предсказания РС.

6. Модель РС «гармоники + шум».

7. Модель сигнала $x[i] = c\ x[i-1] + b\ x[i-2] + n[i]$, где $n[i]$ – белый шум. Как по наблюдениям $\{x[i]\}$ определить значения коэффициентов c, b?

8. Как отличить речевой сигнал от стационарного шума?

9. Частота дискретизации речевого сигнала равна Fs. Какую размерность спектра выбрать, чтобы наблюдать ОТ в узкополосной спектрограмме?

10. Частота дискретизации речевого сигнала равна Fs. Какую размерность спектра выбрать, чтобы наблюдать форманты в широкополосной спектрограмме?

Глава 5
Качество и разборчивость речевых сигналов

Одной из важнейших областей анализа речевых сигналов является оценка их слухового восприятия. Оценка восприятия РС важна в целом ряде областей:

– передача речи по каналам связи;
– сжатие (кодирование) РС;
– распознавание речи;
– идентификация дикторов;
– протезирование и восстановление слуха;
– компенсация шумов РС.

Восприятие речи неразрывно связано, с одной стороны, со свойствами слуха, с другой – свойствами самих речевых сигналов.

Классической книгой по данной области остается книга Фланагана [1]. В данной главе кратко рассмотрены вопросы оценки качества и разборчивости РС.

5.1 Слуховое восприятие звука

Интенсивность и частота являются основными характеристиками звука. Рассмотрим свойства слуха с точки зрения восприятия интенсивности и частоты.

5.1.1 Восприятие интенсивности звука

Интенсивность (сила) звука – средняя по времени энергия, переносимая звуковой волной через единичную площадку, перпендикулярную к направлению распространения волны в единицу времени. Другими словами, интенсивность звука — скалярная физическая величина, характеризующая мощность, переносимую звуковой волной в направлении распространения. Сила звука измеряется в Вт/м².

Звуковое давление – звуковая энергия, которая попадает на единицу площади, расположенную в заданном направлении от источника звука и удаленную от него на определенное расстояние.

Звуковое давление измеряется в паскалях (Па) или микропаскалях (мкПа, 10^{-6} Па).

Мгновенное значение звукового давления в точке изменяется со временем, поэтому практический интерес представляет среднеквадратичное значение данной величины, связанное с интенсивностью звука приведенными ниже соотношениями.

Связь между силой (интенсивностью) звука и звуковым давлением:

$$P = 6{,}4(I)^{\frac{1}{2}}.$$

Звуковое давление в децибелах (*sound pressure level, SPL*):

$$SPL = P\ \text{дБ} = 20\lg\frac{P}{Po},$$

где Po – звуковое давление на пороге слышимости примерно 20 мкПа (20 10^{-6} Па) (порог слышимости при частоте 1 кГц).

Интенсивность звука в децибелах: I дБ $= 20\lg\frac{I}{Io}$,

где I – измеренная интенсивность;

Io – эталонная интенсивность звука (порог при частоте 1 кГц) примерно 10–12 Вт/м².

В шкале децибел интенсивность звука совпадает со звуковым давлением.

Громкость звука – субъективная величина слухового ощущения, которая зависит от интенсивности звука и его частоты. При неизменной частоте громкость звука растет с увеличением интенсивности. При одинаковой интенсивности наибольшей громкостью обладают звуки в диапазоне частот 700–6000 Гц. Нулевой уровень громкости звука соответствует звуковому давлению 20 мкПа и силе звука 10–12 Вт/м^2 на частоте 1 кГц.

Абсолютный уровень громкости принято оценивать в единицах – *сонах*. Громкость в 1 сон – громкость чистого тона с частотой 1 кГц и уровнем звукового давления 40 дБ.

Относительный уровень громкости принято оценивать в логарифмических единицах – *фонах*.

Уровень громкости чистого тона с частотой 1000 Гц в фонах численно равен уровню звукового давления в децибелах.

5.1.2 Область человеческого слуха

Частотная неравномерность слуха описывается соотношением между слышимой громкостью звука и его интенсивностью:

$$L(f) = W(f, I(f)) \times I(f),$$

где $L(f)$ – громкость звука;

$I(f)$ – интенсивность звука;

W – функция, аппроксимирующая неравномерность чувствительности человеческого слуха на различных частотах и интенсивностях звука.

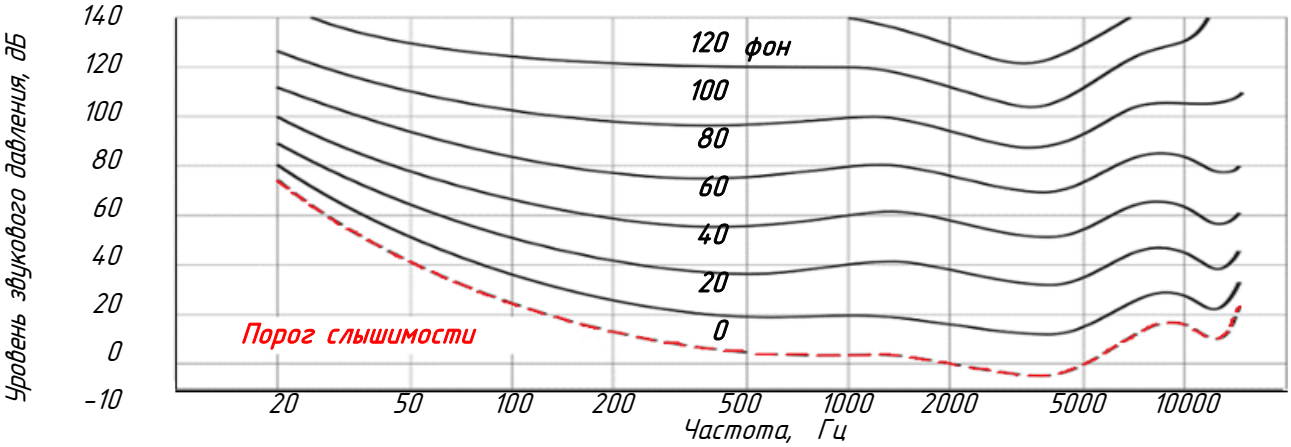


Рисунок 5.1 — Зависимость субъективной громкости звука от звукового давления (кривые равной громкости)

На рисунке 5.1 представлены зависимость относительного уровня громкости от уровня звукового давления (интенсивности звука). Кривые равной громкости стандартизированы [R1].

5.1.3 Восприятие частоты звука

Важнейшим свойством слуха является неравномерная разрешающая способность по частоте. Модели восприятия частоты звука связывают *объективное значение высоты звука* и *субъективное восприятие*.

Для описания частотной селективности человеческого слуха были предложены три шкалы: шкала эквивалентных прямоугольных полос пропускания, шкала барков, шкала мел, характеризующие неравномерность слуха с нескольких разных

позиций [2]. Для каждой из шкал описаны прямые и обратные преобразования с линейной шкалой частот в герцах.

Моделирование слухового восприятия, то есть преобразование между шкалами, можно выполнить двумя способами: непосредственно в пространстве сигналов, применяя банки фильтров, либо в пространстве спектров, суммируя амплитуды спектра в соответствующие группы.

Шкала эквивалентных прямоугольных полос пропускания

Частотная селективность человеческого слуха может быть описана в терминах *эквивалентных прямоугольных полос пропускания* (*equivalent rectangular bandwidth, ERB*) на основе концепции, предложенной Муром [3]. Эквивалентная прямоугольная полоса пропускания представляет собой упрощенную аппроксимацию человеческого слуха идеальными прямоугольными полосовыми фильтрами, полученными на основе экспериментальных исследований. Эквивалентную прямоугольную полосу пропускания можно считать мерой частотного разрешения. Зависимость ширины полосы от частоты приблизительно линейная:

$$b \approx 0{,}108\, f + 24{,}7,$$

где b – ширина полосы фильтра; f – частота.

Часто эту шкалу описывают фильтрами постоянной добротности, коэффициент добротности которых $Q = f/b \approx 9{,}26$.

Шкала барков

Ухо обладает интегрирующей способностью звуков в частотных интервалах. Частоты сложного звука в пределах определенной полосы частот не могут быть индивидуально идентифицированы. В случае же, когда одна из компонент сложного звука выпадает из полосы, она может быть идентифицирована.

Разделение звука на отдельные различаемые группы характеризуется шкалой барков. Барк — психофизическая единица высоты звука, предложенная Э. Цвикером [4]. Эксперименты показали, что в диапазоне до 16 кГц число таких интервалов равно 24. В 1957 г. Цвикером была предложена ставшая канонической концепция критических полос – деление частотного диапазона на критические полосы слуха [4].

Предложенная шкала барков (*bark*) делит частотный диапазон от 20 Гц до 15,5 кГц на 24 критические полосы (*bark filter bank*), границы которых определены экспериментально и представлены в табличной форме:

fn = {100, 200, …12000, 15500 Гц}, n=1, 2, …23, 24.

Предложены аналитические уравнения для перехода между частотами и шкалой барков, аппроксимирующие табулированные значения.

Графическое представление шкалы барков) приведено на рисунке 5.2.

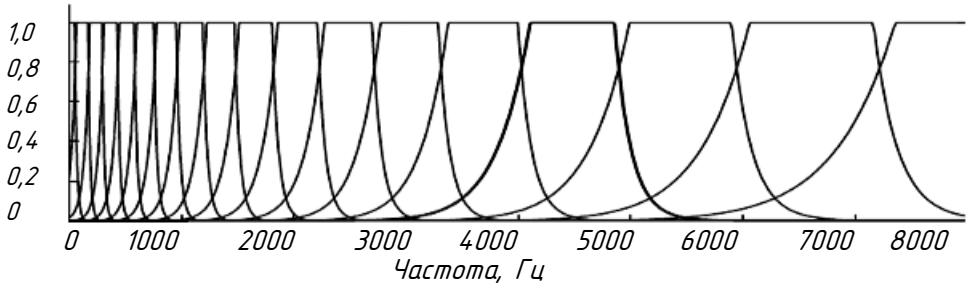


Рисунок 5.2 — Критические полосы фильтров на основе шкалы барков [5]

Шкала мел

Шкала мел представляет собой альтернативное отображение частоты в герцах на шкалу перцептивно значимых частот. Шкала мел – это соотношение между субъективной гармоникой и частотой чистого тона. Шкала мел получила широкое распространение в различных областях анализа и обработки аудиосигналов: для анализа музыкальных сигналов (термин «мел» происходит от слова «мелодия»), задачах детектирования акустических событий, для анализа речевых сигналов.

Мел – это единица измерения высоты звука, основанная на психофизическом восприятии звука человеком и построенная на основе анализа большого числа статистических данных.

Равным дистанциям в мел-шкале соответствуют одинаковые разницы высоты звука, оцениваемые слушателями. Соотношение между шкалой мел и частотой герц описывается соотношением [6]:

$$\text{мел} = 1127 \ln[1 + f/700] = 2595 \lg(1 + f/700),$$

где f – частота в герцах.

Мел аппроксимируется как линейная шкала в диапазоне от 0 до 1000 Гц, а затем в более высоком диапазоне частот как логарифмическая:

$$f \ll 700 \text{ Гц} \rightarrow \text{мел} \approx f \times 1127/700 \approx 1{,}6 \times f,$$
$$f = 1000 \text{ Гц} \rightarrow \text{мел} \approx 1000,$$
$$f > 1000 \text{ Гц} \rightarrow \text{мел} = 1127 \ln[f/700].$$

Пример

Частоте f = 8000 Гц соответствует 2 745 мел

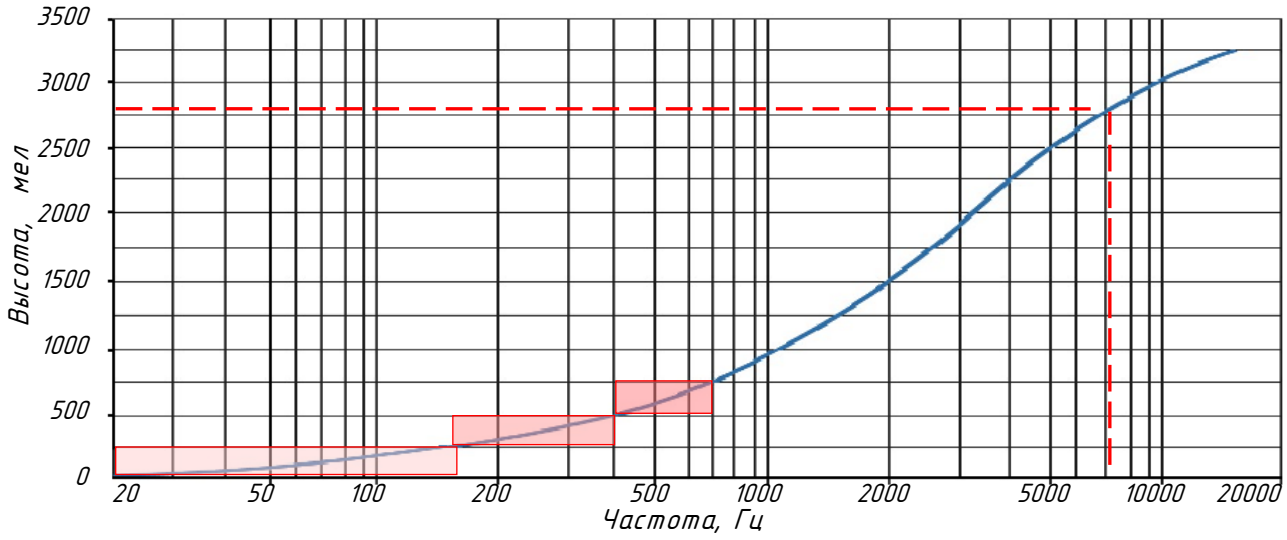


Рисунок 5.3 — График соотношения мел и частоты в герцах [7, 8]

Таким образом, мел-шкала объединяет линейную шкалу в герцах в диапазоне нижних частот и логарифмическую шкалу в герцах в диапазоне высоких частот.

При этом полосам мелов одинаковой ширины соответствуют неравные полосы в линейной шкале частот в герцах и обратно.

Соотношение между шкалами мел и герц иллюстрируется на рисунке 5.3. Между шкалами барков и мел имеется следующее соотношение: 1 барк = 100 мел [9].

Дискретный мел-спектр

Рассмотрим переход от дискретной мел-шкалы к шкале в герцах. Первоначально результатом дискретного преобразования Фурье является дискретный спектр в шкале герц. Переход от равных полос мел-шкалы к неравным полосам шкалы в герцах осуществляется на основе обратного преобразования:

$$f\ \text{Гц}\ =\ 700\ \times \left(\exp\left(\frac{\text{мел}}{1127}\right) - 1\right).$$

Зададимся шагом в мел-шкале. Допустим, в области НЧ он совпадает с шагом линейного спектра. Например, для сигнала 8 кГц и окна анализа 512 шаг и ширина полос равны 8000/512 ≈ 16 Гц). Первой полосе дискретной шкалы в герцах соответствует первая полоса дискретной шкалы мел (1,6×16≈25 мел) и обратно. Второй полосе – вторая. И так приблизительно до 1000 Гц. А дальше полоса шкалы мелов соответствует все большему числу полос шкалы в герцах. Мощности спектра в этих полосах складываются и составляют мощность соответствующей полосы мел. Весь диапазон спектра 8 кГц занимает приблизительно 2700 мел. В дискретной мел-шкале с шагом 25 мел диапазон 8 кГц будет представлен 2700/25=108 полосами мел вместо 512 полос шкалы в герцах. Таким образом, при переходе к мел-шкале достигается значительное сокращение размерности дискретного спектра.

Мел-частотные коэффициенты кепстра

Поскольку ширина мел-полос выбирается в зависимости от решаемой задачи, то возможно более значительное уменьшение их числа. В задачах анализа речи наибольшее распространение получила шкала для вычисления мел-частотных коэффициентов кепстра (*mel-frequency cepstral coefficients, MFCC*).

Основная причина использования мел-шкалы для анализа речи – гипотеза о том, что человеческий слух в лучшей степени приспособлен к детектированию и распознаванию речи. А значит, данные, преобразованные таким образом, будут больше пригодны для автоматического анализа РС. Цель заключается в моделировании неравномерной по частоте избирательности слухового анализатора.

В основу частотного анализа положено разделение шкалы мел на ряд полос равной ширины. Выбор ширины мел-полос определяется практическими соображениями, то есть ширина полос в шкале мел не является фиксированной, а зависит от числа выбранных мел полос.

Для задачи распознавания речи используют 12–20 мел полос. На рисунке 5.4 а) шаг и полуширина полос составляет 200 мел.

При вычислении коэффициентов *MFCC* мощности полосы в герцах объединяются в общую мощность широкой полосы. Это достигается объединением амплитуд линейного спектра мощности в группы разной ширины частот, соответствующие группам равной ширины в шкале мел (рисунок 5.4 б)).

Объединение амплитуд линейного спектра мощности в группы осуществляется с помощью частотных фильтров треугольной формы. Учет различной ширины полос мел-спектра в амплитудах осуществляется непосредственно в амплитудах фильтра либо в дальнейшем за счет умножения амплитуд мел-спектра на соответствующие веса. Далее амплитуды мел-спектра логарифмируются, после чего к логарифмам амплитуд применяется дискретное косинусное преобразование (*discrete cosine transformation DCT*).

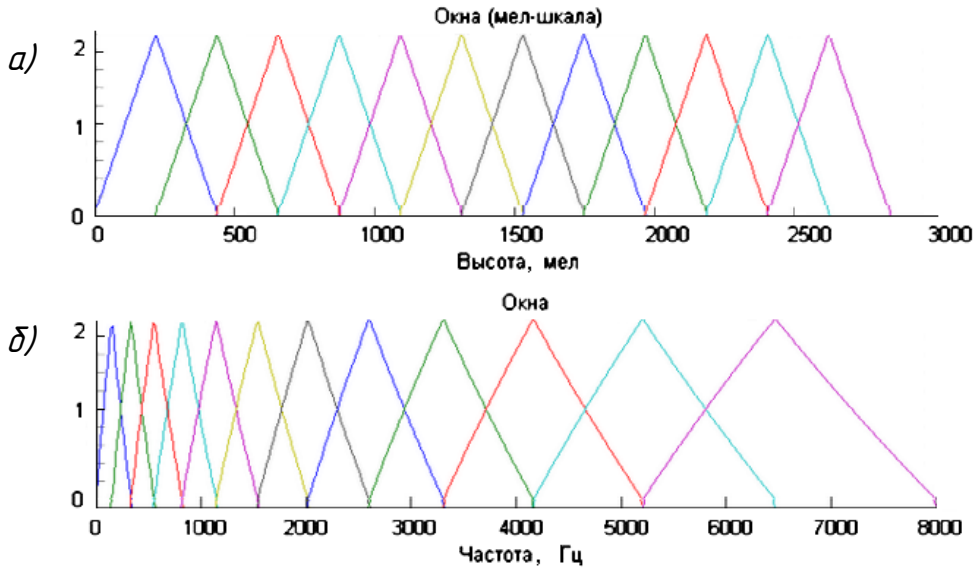


Рисунок 5.4 — Полосы фильтров *MFCC*: а) мел-шкала, б) шкала в герцах [9]

Общая формула вычисления коэффициентов *MFCC*:

$$MFCC(n) = DCT\{\lg(\mathbf{M} \times \mathbf{P}xx)\},$$

где **M** – матрица весовых коэффициентов для фильтров в мел шкале;

P*xx* – вектор дискретного спектра мощности;

DCT – дискретное косинусное преобразование.

5.2 Качество и разборчивость речевых сигналов

Речь — это в первую очередь способ передачи мысли (информации). Речь может также быть направлена на передачу эмоций. Помимо лингвистической информации и эмоций, речь содержит много другой информации (личность говорящего, физические состояние и пр.). Поэтому свойства РС можно характеризовать со многих сторон. С позиции передачи информации выделяют разборчивость РС. Совокупность различных характеристик можно отнести к качеству РС. Понятие качества применяется не только для РС, но и для характеристики других аудиосигналов. Например, для музыкальных сигналов применяют десятки различных характеристик.

Качество речевых сигналов отражает комфортность восприятия РС конкретным слушателем, ясность, искажения и уровень шума в сигнале.

Разборчивость речевых сигналов является способностью слушателя расшифровывать содержимое речи. Под разборчивостью речи, прежде всего, понимается возможность извлечения информации, передаваемой в речевом сообщении. Мерой разборчивости РС является доля правильно распознанных элементов речи (фонемная, слоговая, словесная и фразовая разборчивость).

Оценка качества и разборчивости РС важна для решения основных задач обработки речевых сигналов:

– передача речи по каналам связи;
– сжатие (кодирование) РС;
– распознавание речи;
– идентификация дикторов;
– протезирование и восстановление слуха;
– компенсация шумов РС.

Разделение критериев качества и разборчивости в некоторой степени условно, поскольку качество речевых сигналов может влиять на его разборчивость и обратно.

Качество можно рассматривать как более широкую категорию, включающую в себя и разборчивость. Тем не менее, к настоящему моменту сложилась система отдельных оценок качества и разборчивости речевых сигналов.

5.2.1 Факторы качества и разборчивости

На качество и разборчивость речи влияет большое число факторов. Перечислим некоторые из них.

Субъективные факторы:

– потеря слуха;

– усталость слушателя (длительное прослушивание сигнала плохого качества);

– акцентная и шепотная речь;

– тихая речь;

– непрерывная речь, голосовой коктейль, ломбард-эффект;

– спектральная и временная маскировка.

Объективные факторы:

– большая дистанция между диктором и микрофоном;

– неправильные режимы записи (недостаточный уровень сигнала; перегрузки, частота дискретизации, разрядность квантования и др.);

– плохая аппаратура (нелинейные искажения, наводки);

– высокий уровень шума окружения;

– эхо и реверберация;

– шумы каналов передачи сигнала;

– способ кодирования информации;

– искажения в каналах связи (малая полоса пропускания, потеря пакетов и др.).

Влияние каждого из факторов на разборчивость речи является предметом отдельного исследования. Затронем кратко лишь один аспект разборчивости: насколько различные частотные диапазоны важны для кодирования и понимания речи.

Прежде всего, разборчивость РС связана со спектром речевых сигналов.

На рисунке 5.5 приведены условные графики среднего спектра русской речи и график вклада в разборчивость областей спектра речи.

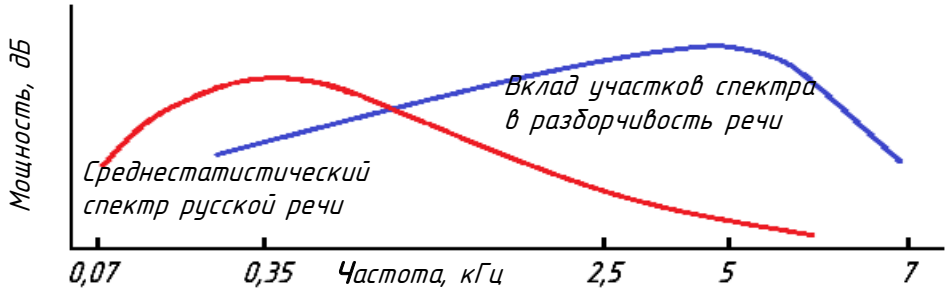


Рисунок 5.5 — Вклад в разборчивость областей спектра речи [10]

Численные соотношения между энергией участков спектра и вкладом в разборчивость представлены ниже.

Важными частотами для разборчивости речи являются звуки согласных в диапазоне от 500 до 4000 Гц. Они определяют 83 % разборчивости.

Диапазон частот от 100 до 3400 Гц обеспечивает более 90 % разборчивости речи.

Частоты от 500 до 1000 Гц дают 35 % словесной разборчивости и содержат 35% звуковой энергии.

Частоты от 1000 до 4000 Гц дают 48 % разборчивости и содержат только 4% звуковой энергии.

Частоты от 125 до 500 Гц содержат 55 % звуковой энергии и дают лишь 4 % разборчивости.

Звуки согласных играют важную роль для разборчивости, несмотря на то, что их энергия намного меньше энергии гласных.

Основная часть энергии генерируется на тональных звуках и концентрируется в 3–4 формантах вплоть до частоты 2 кГц.

Как было отмечено в п. 5.2, чувствительность слуховой системы к различным частотам различна. Из всего диапазона 20–20000 Гц самые низкие пороги восприятия (максимальная чувствительность слуха) относятся к диапазону 2–5 кГц, что приблизительно соответствует области максимального вклада участков спектра в разборчивость речи.

Области спектра с большой звуковой энергий РС не совпадают с областью, имеющей наибольшее значение для разборчивости. Данный факт может оказаться важным в решении других задач.

Например, вклад различных спектральных составляющих в информацию об акустическом событии зависит от характера акустических процессов и не совпадает со спектральным распределением вклада в разборчивость речи.

Изучение информативности различных участков спектра остается важной практической задачей во многих областях.

5.2.2 Оценка качества и разборчивости

Оценка качества и разборчивости важны во многих областях. Для различных приложений предложены различные меры их количественной оценки. Эти меры можно разделить на две большие группы: субъективные и объективные.

Субъективные оценки основаны на прослушивании и оценке аудиоматериала «наивными» слушателями или профессиональными слушателями (аудиторами).

Объективные оценки предполагают автоматизированный расчет различных числовых мер и предназначены для прогнозирования результатов субъективной оценки. Субъективные оценки трудозатратны, так как требуют участия специалистов – дикторов и аудиторов, но достоверны, объективные оценки – менее надежны.

Для учета различных типов искажений РС разработаны различные объективные критерии, многие из которых описаны в соответствующих стандартах. Как правило, алгоритмы расчета стандартизованных оценок реализованы в программах, поставляемых вместе со стандартами. Использование таких программ требует определенной квалификации.

Субъективные и объективные оценки качества и разборчивости также можно разделить по другому признаку: на те, что основаны на сравнении тестового сигнала

с эталоном и оценки, выполняемые без эталона. Меры, основанные на сравнении образцового РС с тестируемым, называются *относительными*, меры, не использующие образцовый сигнал – *абсолютными*, а методы их оценки соответственно – *интрузивными* и *неинтрузивными*.

Абсолютные субъективные оценки основаны на опыте аудиторов, абсолютные объективные меры основаны на предварительных оценках корреляции между параметрами сигнала и его качеством и разборчивостью.

Подавляющая часть объективных мер основана на сравнении сигналов с эталоном. Однако в последние годы предпринимаются попытки разработки объективных мер, не использующих образцовый сигнал.

Общая классификация методов оценки качества и разборчивости РС приведена на рисунке 5.6.

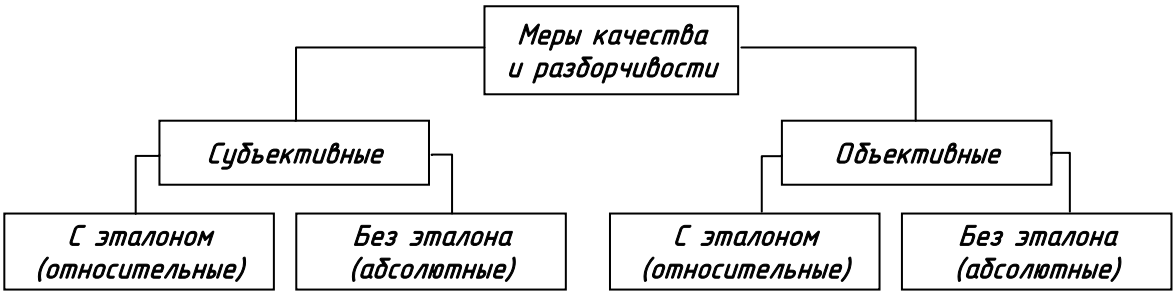


Рисунок 5.6 — Классификация методов оценки качества и разборчивости РС

Рассмотрим основные группы мер качества и разборчивости. Подробную информацию можно найти в работах [11, 12].

5.3 Методы оценки качества речевых сигналов

Объективные оценки качества обычно основаны на сравнении исходного (эталонного) и оцениваемого РС.

Субъективные оценки качества также могут основываться на сравнении оцениваемого сигнала с эталоном, но некоторые оценки основываются на субъективном мнении (и опыте) аудиторов.

Основные меры качества (субъективные и объективные оценки) перечислены в таблице 5.1.

Таблица 5.1 — Меры оценки качества речевых сигналов

Субъективные оценки качества	Объективные оценки качества
MOS (mean opinion score)	Временные: *SNR, SNRseg, fwSNRseg*
DMOS (degradation mean opinion score)	Спектральные: *SD, ISD, LSD, LAR*
PPT (pair-wise preference test)	Перцептивные: *SNA, PSQM, PESQ, BSD*
DAM (diagnostic acceptability measure)	

Рассмотрим некоторые из них.

5.3.1 Субъективное оценивание качества

Наиболее известным субъективным способом оценки качества речевых сигналов является средняя экспертная оценка *MOS* (*mean opinion score*) — субъективно-статистические испытания с помощью группы слушателей-экспертов.

Средняя экспертная оценка вычисляется как среднее арифметическое от *N* независимых экспертных оценок *Rn*:

$$MOS = \frac{1}{N}\sum_{n=1}^{N} R_n.$$

Используются два метода оценки *MOS* – абсолютная и относительная (дифференциальная средняя экспертная оценка) *DMOS*.

В методе абсолютной оценки (*absolute category rating, ACR*) слушатели сравнивают искаженный сигнал с их собственной внутренней моделью речи хорошего качества.

В методе относительной оценки (*degradation category rating, DCR*) слушатели вначале прослушивают неискаженную речь, а после этого оценивают ухудшение (качество) искаженной речи.

Методы обеих оценок определены в стандарте ITU [E6] и отечественных стандартах [R3, R5].

Для разных задач используют разные пятибальные шкалы оценок. Три наиболее используемых в исследованиях шкалы мнений: *шкала слухового качества звучания, шкала усилий при прослушивании, шкала ухудшения качества.*

В таблице 5.2 приведены первые два варианта шкал оценки качества РС.

Таблица 5.2 — Шкалы оценок качества MOS и DMOS

Оценка	Шкала качества звучания	
	Абсолютное качество звучания ACR, MOS	Степень ухудшения качества звучания DCR, MOS (DMOS)
5	Прекрасно	Незаметно
4	Хорошо	Заметно, но не вызывает неприятных ощущений
3	Удовлетворительно	Немного вызывает неприятное ощущение
2	Неудовлетворительно	Вызывает неприятное ощущение
1	Плохо	Вызывает очень неприятное ощущение

Рассмотренные методики могут быть приняты за основу оценки других характеристик речевых сигналов, например в задачах синтеза речи и шумоочистки аудиосигналов.

5.3.2 Объективные оценки качества РС

Объективные оценки качества РС основаны на сопоставлении образцового и искаженного сигналов. Такие методы называются эталонными. Результатом такого сравнения, как правило, является скалярная величина, характеризующая дистанцию между образцовым и тестовым сигналами в некотором измерении. Основные группы оценок перечислены ниже.

Во временном измерении:
SNR (signal to noise ratio)
SNRseg (segmental SNR)

В спектральном измерении:
FWSegSNR (frequency weighted segmental SNR)
SD (spectral distance)
ISD (itakura-saito distortion)
LSD (log spectral distance)
LAR (log-area ratio)
LLR (log likelihood ratio) for word features evaluation (LPC-based measure)

В кепстральном измерении:
CD (cepstral distance)

В аудитивном измерении (*perceptual motivated measures*):
WSS (weighted spectral slope measure)
SNA (A-weighted signal/noise ratio)
PSQM (perceptual speech quality measure)
PESQ (perceptual evaluation of speech quality)
BSD (bark spectral distortion)

Обзор этих мер дан в работах [11, 12]. В качестве примера рассмотрим меру, основанную на оценке отношения сигнал-шум.

Общее отношение сигнал-шум, вычисленное для всего речевого сигнала, имеет низкую корреляцию с качеством, поэтому редко применяется в качестве объективного показателя.

Бóльшее распространение получило сегментное ОСШ (*segSNR*), являющееся одной из простейших мер. Для правильного вычисления этой характеристики необходимо точное выравнивание сигналов на временной оси. Сегментное ОСШ во временной области определяется следующим образом:

$$segSNR \text{ дБ} = 1/K \sum_{k=1}^{K} SNR(k) \text{ дБ} = < 10 \lg[Ps(k)/Pn(k)] >_{sp},$$

где k – индекс кадров речи;

K – количество кадров речи;

$< \cdots >_{sp} = 1/\text{K} \sum_{k=1}^{K}$ – знак усреднения по кадрам речи длительностью 15–20 мс;

$Ps(k)$ и $Pn(k)$ – мощности сигнала и шума (искажения) на кадрах речи.

Сегментное ОСШ может также вычисляться в частотной области. В этом случае мощности сигнала и шума вычисляются через спектры.

Наконец, в спектральном измерении используют частотно-взвешенную оценка ОСШ (*FWSegSNR, frequency weighted segmental SNR*), учитывающую различный вклад частотных полос в слуховое восприятие человека.

Перечисленные меры объективной оценки качества являются интрузивными, то есть основаны на сравнении анализируемого сигнала с эталоном.

Неинтрузивный (пассивный) метод описан в стандарте P563 [E4].

5.4 Методы оценки разборчивости речевых сигналов

Методы оценки разборчивости речи можно, прежде всего, разделить на субъективные и объективные. К настоящему времени большое количество мер кодифицировано в стандартах.

5.4.1 Методы субъективной оценки разборчивости

Субъективные методы тяжелы в реализации, так как требуют участие специалистов: дикторов и аудиторов. Но так как разборчивость речи - это относительное количество правильно принятых элементов артикуляционных таблиц, а конечной принимающей стороной является исключительно человек, то данные методы считаются наиболее достоверными и используются в составлении баз данных для разработки и тестирования объективных методов оценки.

Числовая мера разборчивости речи – это относительное количество правильно распознанных элементов. Различают *фразовую, словесную и слоговую разборчивость*. Основная идея тестов заключается в сравнении произнесенных и распознанных слов и фраз, собранных в артикуляционных таблицах.

Методы субъективной оценки разборчивости определены и подробно изложены в стандартах [R2–R5, E1–E3].

Основные меры субъективной разборчивости РС перечислены в таблице 5.3.

Таблица 5.3 — Меры субъективной разборчивости речи

Оценки субъективной разборчивости	
DRT (diagnostic rhyme test)	Диагностический тест на рифму
MRT (modified rhyme test)	Модифицированный тест на рифму
SUS (semantically unpredictable sentences)	Семантически непредсказуемые предложения
DALT (diagnostic alteration test)	Диагностический тест на изменение
DMCT (diagnostic medial consonant test)	Диагностический тест на медиальный согласный

5.4.2 Методы объективной оценки разборчивости

Методы объективной оценки можно разделить на две группы.

Первая группа - это методы, использующие эталонный сигнал (эталонные методы). Данные методы применимы для анализа систем, свойства которых не будут существенно меняться со временем, например, для оценки акустики помещений.

Вторая группа - это методы, способные делать оценку разборчивости непосредственно по самому тестируемому сигналу.

Эталонные методы оценки разборчивости речи определяют факторы и их влияние путем сравнения чистого (эталонного) и искаженного сигнала. В первом приближении выделяют два фактора: значимости частотных полос и степень амплитудной модуляции звуков в этих полосах, характеризуемой отношением сигнал-шум.

Соответственно методы оценки разборчивости можно разделить на две группы: *формантные* и *модуляционные*.

Формантные методы основаны на оценке ОСШ в определенных полосах частот [13]. Идея данных методов оценки заключается в корреляции разборчивости речи с тем, насколько хорошо сохранены форманты речи в зависимости от ОСШ в по-

лосах частот. Основными формантными мерами являются А-взвешенное долговременное ОСШ, индекс артикуляции и индекс разборчивости речи.

Основные меры объективной разборчивости речи перечислены в таблице 5.4.

Таблица 5.4 — Меры объективной оценки разборчивости речи

Формантные	Модуляционные
SNA (A-weighted Signal/Noise Ratio) *AI (Articulation Index)* *SII (Speech Intelligibility Index)*	*STI (Speech Transmission Index)* *RaSTI (Rapid STI)* *STOI (Short-Time Objective Intelligibility)* *STITEL (STI for Telecommunication Systems)* *STIPA (TI for Public Address Systems)*

Рассмотрим основные эталонные методы объективной оценки разборчивости речи.

В оценке разборчивости в первом приближении выделяют два фактора: значимости частотных полос и степень амплитудной модуляции звуков в этих полосах, характеризуемой отношением сигнал-шум. Изучению этих факторов посвящено большое число отечественных и зарубежных исследований.

Методы объективной оценки можно разделить на две группы.

Первая группа - это методы, использующие эталонный сигнал. Данные методы применимы для анализа систем, свойства которых не будут существенно меняться со временем, например, для оценки акустики помещений.

Вторая группа - это методы способные делать оценку разборчивости непосредственно по самому тестируемому сигналу. Эталонные методы оценки разборчивости речи определяют факторы и их влияние путем сравнения чистого (эталонного) и искаженного сигнала.

Рассмотрим основные из них. Наиболее простыми для понимания являются методы, основанные на оценке ОСШ в определенных полосах частот, и по от этим данным дающие оценку разборчивости [13].

Идея данных методов оценки заключается в корреляции разборчивости речи с тем насколько, хорошо сохранены форманты речи в зависимости от ОСШ в полосах частот.

SNA – А-взвешенное долговременное ОСШ

В качестве меры разборчивости принимается *А-взвешенное долговременное ОСШ:*

$$SNA\,дБ = Ps_A\,дБ - Pn_A\,дБ$$

AI – индекс артикуляции

Индекс артикуляции и его различные модификации рассмотрены в работах [11, 12]. Идея метода состоит в разбиении спектра сигнала на некоторое количество полос и измерении ОСШ в каждой из них.

Данный метод основывается на предположении, что каждая полоса вносит свой независимый вклад в общую разборчивость речевого сигнала. Параметры полос выбираются так, чтобы каждая из них имела равный вклад в общую разборчивость.

На первом шаге вычисляются ОСШ в двадцати равноартикуляционных полосах частот. Затем полученные отношения ограничиваются, нормализуются и для них

вычисляется взвешенное среднее:

$$AI \approx 0{,}05 \sum_{k=1}^{20} \min\{1, SNAi\_dB/30\} .$$

В более общей форме разборчивость речи оценивается так:

$$AI = \frac{1}{K} \sum_{k=1}^{K} p_k * P(\Delta L_k) ,$$

где K – количество полос частот;

p_k – вероятность наличия форманты в полосе частот;

$P(\Delta L_k)$ – вероятность, того, что речь не будет скрыта шумом;

ΔL_k – отношение между пиковым уровнем речи и эффективным уровнем шума.

SII Индекс разборчивости речи

SII (speech intelligibility index) – индекс разборчивости речи - был предложен как дальнейшее развитие метода *AI* и включен в американский стандарт [E1]. Метод измерения *SII* описан в [12, R6].

Этот метод подразумевает измерение ОСШ в отдельных полосах с последующим суммированием значений, характеризующих разборчивость в отдельных полосах. Главное отличие от критерия *AI* состоит в том, что вместо набора из 20 полос в данном подходе используются 4 различных набора полос:

– критические полосы;

– третьоктавные полосы (21 полоса);

– критические полосы, обладающие равным вкладом в разборчивость (21 полоса);

– октавные полосы (7 полос).

Значение критерия *SII* изменяется от 0 до 1.

Модуляционные

– индекс передачи речи (*speech transmission index, STI*);

– индекс передачи речи в телефонных системах (*STI for telecommunication systems, STITEL*);

– индекс объективной кратковременной разборчивости (*short-time objective intelligibility, STOI*);

– быстрый индекс передачи речи (*Rapid STI, RaSTI*);

– индекс передачи речи для системы громкой связи (*TI for Public Address Systems, STIPA*).

STI – индекс передачи речи

Индекс передачи речи *STI* (*speech transmission index*) является объективным методом *прогнозирования и измерения* речевой разборчивости [12, 14–16,].

Методы измерения разборчивости с помощью *STI* введены в международный стандарт [E3]. Метод предполагает проведение измерений в широком диапазоне частот и учитывает частотную зависимость времени реверберации, неравномерность амплитудно-частотной характеристики и другие частотно-зависимые эффекты, что в результате дает достаточно хорошую корреляцию с субъективными оценками. Для оценки *STI* используется тестовый сигнал.

Для прогнозирования разборчивости оценивается уменьшение модуляции интенсивности огибающих сигнала в октавных частотных полосах. Метод *STI* основан на том, что потери в амплитудной модуляции можно считать мерой потери разборчивости. При этом частоты модуляции, как и октавные полосы, выбираются так, чтобы соответствовать частотам естественной человеческой речи. В базовом варианте выбирается семь октавных полос от 125 Гц до 8 кГц и 14 третьоктавных частот модуляции от 0,63 Гц до 12 Гц. Полученные коэффициенты преобразуются в отношения сигнал-шум, находится взвешенное среднее и нормализуется. Схема работы алгоритма показана на рисунке 5.7.

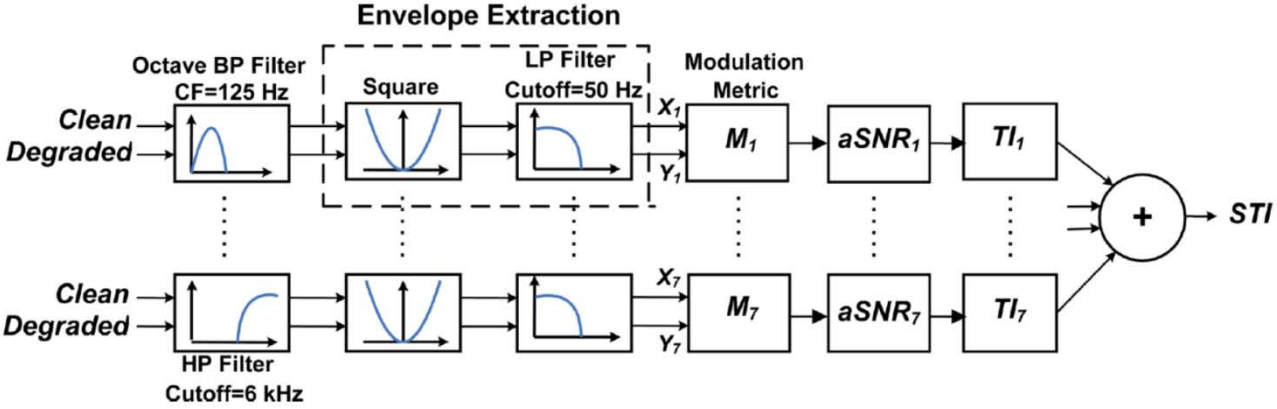


Рисунок 5.7 — Схема измерения значений *STI* [17]

Схема обработки аналогична схеме вычисления модуляционного спектра (п. 2.5.5).

Образцовый и тестовый сигналы разделяются на полосовые сигналы октавными фильтрами (*Octave BP Filter*), далее в каждой полосе частот вычисляется огибающая сигналов (*Square*). Для дальнейшего спектрального анализа, выполняемого от кадра к кадру с частотой следования кадров, частота полосы огибающей ограничивается величиной 50 Гц, чтобы избежать эффекта наложения спектров. После этого к текущим огибающим применяется дискретное преобразование Фурье, после чего вычисляется дискретный спектр мощности. Матрица значений спектров мощности в частотных полосах сигнала образует матрицу значений модуляционного спектра (*Modulation Metric*). Модуляционные компоненты спектров образцового и тестового сигналов преобразуются в модуляционные отношения сигнал-шум, по которым вычисляются индексы передачи (*TI*), которые объединяются в общий индекс передачи речи (*STI*).

У данного метода существует несколько модификаций. Часть из них направлена на уменьшение количества требуемых измерений, например *RASTI* [18], *STITEL*.

Основная идея этих модификаций заключается в сужении области применения алгоритма и соответственно оптимизации его под конкретную задачу.

Метод *RASTI* использует октавные полосы, равные только 500 и 2000 Гц, а также уменьшает количество частот модуляции: 4 для октавной полосы 500 Гц и 5 для 2 кГц. Метод используется при оценке акустики помещений [16].

Метод *STITEL* применяется при оценке телефонных систем [19].

Индекс объективной кратковременной разборчивости (*short-time objective intelligibility STOI*) является функцией оценки кратковременной корреляции частотно-временных огибающих чистой и анализируемой речи с помощью коэффициента корреляции [20].

Оценка *STOI* состоит из следующих этапов:

1. Разделение образцового и анализируемого входных сигналов на третьоктавные полосовые сигналы с использованием ДПФ.

2. Разбиение полосовых сигналов на кадры и удаления кадров, в которых нет речевого сигнала.

3. Вычисление корреляции между полосовыми образцовым и анализируемым сигналами.

4. Вычисление средней по времени и полосам корреляции образцового и анализируемого и сигналов.

Схема алгоритма представлена на рисунке 5.8.

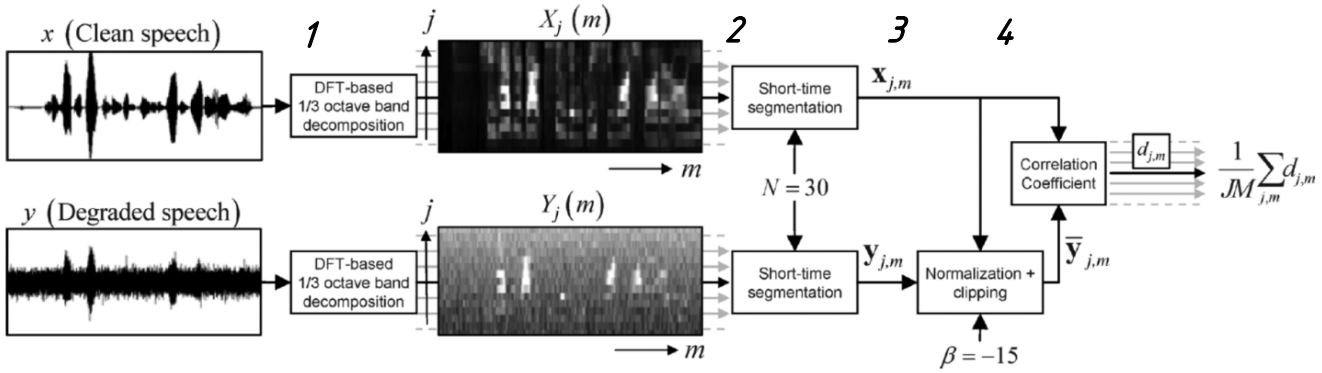


Рисунок 5.8 — Схема алгоритма вычисления метрики *STOI* [16]

В заключение отметим, что задача оценки качества и разборчивости без предъявления тестового сигнала по-прежнему остается актуальной.

Подходы к её решению рассмотрены в работах [21–25].

Однако они обладают большей сложностью и, как правило, обеспечивают требуемый уровень достоверности лишь для ограниченного класса искажений или внешних условий.

Литература

1. Фланаган Дж. Анализ, синтез и восприятие речи. — М.: Связь, 1968.

2. Гуртуева И.А., Бжихатлов К.Ч. Аналитический обзор и классификация методов выделения признаков акустического сигнала в речевых системах // Известия Кабардино-Балкарского научного центра РАН. 2022. No 1 (105). С. 41–58.

3. Moore B. Frequency selectivity in hearing. Boston, MA: Springer, 1986. P. 456.

4. Цвикер Э., Фельдкеллер Р. Ухо как приемник информации. — М.: Сов. Радио, 1971. — 256 с.

51. Gajsek R. and Mihelic F. Comparison of speech parameterization techniques for Slovenian language // 9th International PhD Workshop on Systems and Control: Young Generation Viewpoint 2008, Slovenia.

6. Алдошина И., Приттс Р. Музыкальная акустика. Учебник // СПб.: Композитор, 2006, 720 с. ISBN 978-5-73790-298-8

7. Rabiner L. R. and Schafer R. W. Introduction to Digital Speech Processing // Foundations and Trends in Signal Processing. Vol. 1, No.1–2 (2007) 194 P.

8. Introduction to Speech Processing https://speechprocessingbook.aalto.fi/

9. Lerch A. An introduction to Audio Content Analysis // John Wiley & Sons, Inc., 2012.

10. Общие положения по специальным исследованиям акустического и виброакустического канала https://ppt-online.org/663428

11. Taal C.H. et al. A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech, ICASSP 2010, pp/3013-3027.– 2010.

12. Топников, А. И. Цифровая обработка речевых сигналов: практикум / Яросл. гос. ун-т им. П. Г. Демидова – Ярославль: ЯрГУ, 2018. – 40 с. – 2011.

13. Тампель И.Б., Карпов А.А. Автоматическое распознавание речи. Учебное пособие. − СПб: Университет ИТМО, 2016. – 138 с.

14. Steeneken H. Development of an Accurate, Handheld, Simple-to-use Meter for the Prediction of Speech Intelligibility. 2001.

15. Galindo, M., Zamarreño, T., Girón, S. Comparative study of various techniques to measure speech intelligibility, 2002. – 2002

16. Cees H. Taal et al. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech – IEEE Transactions on Audio, Speech, and Language Processing – vol 19, № 7 – 2011

17. Payton K. L and Shrestha M. Comparison of a short-time speech-based intelligibility metric to the speech transmission index and intelligibility data // The journal of the Acoustical Society of America – 134, 3818 – 2013.

18. Ma J. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions // The Journal of the Acoustical Society of America – 125, 3387 – 2009.

19. Falk T. H., Zheng C., Wai-Yip Chan. A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech – IEEE Transactions on Audio, Speech, and Language Processing – vol. 18. – 2010.

20. Tiago H. Falk, Chenxi Zheng, Wai-Yip Chan A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech – IEEE Transactions on Audio, Speech, and Language Processing – vol. 18, issue 7 – 2010

21. V. Grancharov V. et al. Low-Complexity, Nonintrusive Speech Quality Assessment // IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 1948–1956, 2006.

22. Sharma D/ et al. Data driven method for non-intrusive speech intelligibility estimation – 2010 18th European Signal Processing Conference – 2010

23. Leman A, Faure J., Parizet E. A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises

24. Ragano A; Benetos E, and Hines A. More for Less: Non-Intrusive Speech Quality Assessment with Limited Annotations // 13-th Int. Conf. on Quality of Multimedia Experience, 2021.

Стандарты

R1. ГОСТ Р ИСО 226-2009 Акустика. Стандартные кривые равной громкости. – М.: Стандартинформ, 2010.

R2. ГОСТ 16600-72 Передача речи по трактам радиотелефонной связи. Требования к разборчивости речи и методы артикуляционных измерений. – Межгосударственный стандарт, 1974. – 74 с.

R3. ГОСТ Р 50840-95 Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. – М.: Госстандарт России, 1997. – 234 с.

R4. ГОСТ 25902-83. Зрительные залы. Методы определения разборчивости речи. – М.: Государственный комитет по делам строительства, 1983.

R5. ГОСТ Р 51061-97 Системы низкоскоростной передачи речи по цифровым каналам. – М.: Госстандарт России, 1997. – 24 с.

R6. ГОСТ Р ИСО 9921—2013 Эргономика. ОЦЕНКА РЕЧЕВОЙ СВЯЗИ Приложение С. Индекс передачи речи. – М.: Стандартинформ, 2014.

E1. ANSI Standard S3.5-1997 (R2007). Methods for Calculation of the Speech Intelligibility Index, 1997.

E2. ISO/TR4870: 1991. Acoustics – The construction and calibration of speech intelligibility tests

E3. IEC 268-16. 2020 Standard. Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index.

E4. ITU-T P.563 Single-ended method for objective speech quality assessment in narrow-band telephony applications – 2004

E5. Methods for subjective determination of transmission quality. ITU-T Recommendation P.800, Aug. 1996.

E6. ITU-T Rec.P.800, Absolute category rating (ACR) method. Annex B, 1996.

E7. ITU-T Rec. P. 862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

Вопросы и задачи

1. Во сколько раз уровень звукового давления неприятных (болевых) ощущения звука превосходит минимальный уровень слышимого звука?

2. Механизм слуха. Область звукового восприятия (диапазон человеческого слуха).

3. От чего зависит разборчивость речи? Какие меры разборчивости речи вам известны? Как оценить разборчивость речи, записанной на фонограмме?

4. Что понимается под качеством речевого сигнала? Как оценить качество разборчивость речи, записанной на фонограмме?

5. В чем отличие между субъективными и объективными методами оценки качества?

6. Назовите наиболее распространенные объективные методы оценки качества речевых сигналов.

7. Что понимается под разборчивостью речевого сигнала? В чем состоит отличие между разборчивостью и качеством?

8. Как производится субъективная оценка разборчивости речи?

9. Как производится объективная оценка оценки разборчивости?

10. Назовите наиболее распространенные формантные методы оценки разборчивости. В чем состоит общий принцип их работы?

11. В чем состоит основная идея модуляционных методов оценки разборчивости? Какой сигнал используется для измерений?

12. Запишите центральные частоты семи октавных полос речи: 125 Гц, …

Заключение

Круг задач анализа и обработки акустических сигналов активно расширяется на протяжении последних десятилетий. Переход от аналоговой обработки к цифровой и далее к нейронным сетям открыл в области речевых технологий принципиально новые возможности и сделал актуальным решение новых задач. Материал данного пособия является основой для освоения современных технологий цифрового анализа и обработки речевых и акустических сигналов.

Важным направлением в области речевых технологий в последнее время становится задача извлечения речевой информации с использованием сигналов удаленных микрофонов, в том числе с использованием микрофонных решеток. В классически курсах по обработке речевых сигналов эти вопросы практически не освещены. При выборе излагаемых в книге материалов мы имели в виду дальнейшее рассмотрение этих вопросов.

Еще один аспект заключается в следующем. Речевые сигналы помимо непосредственного содержания несут информацию об условиях распространения звуков, характеристиках помещений, устройствах и особенностях звукозаписи. В связи с этим актуальной является задача извлечения из акустических сигналов содержащейся в них разнообразной информации. К этому же кругу задач относятся методы детектирования акустических событий, распознавание акустических сцен, распознавания эмоций, активно обсуждаемые в настоящее время. Рассматриваемые в книге характеристики также выбраны с учетом дальнейшего рассмотрения этих вопросов.

Автор последовательно на протяжении всей книги придерживается мысли, что «Цель расчетов – понимание, а не числа» (Р. Хэмминг, «Численные методы для научных работников и инженеров» – М.: Наука, 1972 г.). Понимание – это особенность человека.

Приложение А
Список сокращений

АКФ – автоковариационная функция
АРУ – автоматическая регулировка усиления
БПФ – быстрое преобразование Фурье
ВЧ – высокие частоты
ДПФ – дискретное преобразование Фурье
КАС – кратковременный анализ сигналов
КСА – кратковременный спектральный анализ
КЛП – коэффициенты линейного предсказания
МС – модуляционный спектр
НЧ – низкие частоты
ОТ – основной тон
ОСШ – отношение сигнал-шум
РС – речевой сигнал
СА – спектральный анализ
СКЗ – среднеквадратичное значение
ФЛП – фильтр линейного предсказания
ФНЧ – фильтр низких частот
ФВЧ – фильтр высоких частот
ФЭС – фильтр экспоненциального сглаживания
ЧПН – частота пересечений нуля

Приложение Б
Список обозначений и символов

Ak – коэффициенты линейного предсказания (КЛП)
A – вектор КЛП (**A** = $[A_1, A_2, \ldots A_p]^T$
A[*k*] – вектор КЛП на кадре *k*
Ax[*f*] – средний по времени амплитудный спектр сигнала *x*[*i*]
Ax[*n*] – средний по времени дискретный амплитудный спектр сигнала *x*[*i*]
Ax[*n, k*] – дискретный амплитудный спектр сигнала *x*[*i*] на кадре *k*
a_i – рекурсивные коэффициенты фильтра

b – шаг (бин) дискретных отсчетов в области частот (шаг =1/NT) в герцах
b_i – нерекурсивные коэффициенты фильтра

c –скорость звука в воздухе (340 м/с)
Cxx[*m*] – дискретная автоковариационная функция процесса *x*[*n*], соответствующая временному сдвигу *m*
Cxy[*m*] – дискретная кросс-ковариационная функция (взаимная ковариация) двух процессов *x*[*n*] и y[*n*], соответствующая временному сдвигу *m*
Cxy – матрица кросс-ковариаций (ковариационная матрица) двух процессов *x*[*n*] и y[*n*]
Cxy – ковариация двух случайных величин
CCx[*m*] – линейно-частотные коэффициенты кепстра
CCx[*m, k*] – линейно-частотные коэффициенты кепстра на *k*-м кадре

D – дистанция
Dx – дисперсия случайной переменной *x*
DFT{ } – дискретное прямое преобразование Фурье

e – 2,7182818…
$\exp(x) = e^x$
Ex – энергия сигнала
Ex[*k*] – энергия сигнала на *k*-м кадре
Ep[*i*] – огибающие мощности
Ex[*i*] – огибающие сигнала
E{*x*} – математическое ожидание случайной величины *x*
E{*x*[*i*]} – математическое ожидание дискретного сигнала *x*[*i*]

f – частота (циклов, выборок, операций в секунду)
Fs – частота дискретизации сигнала, Гц = 1/T
Fn – частота Найквиста, Гц
Fk – частота следования кадров, Гц
Fm – модуляционная частота, Гц
Fo – частота основного тона, Гц

Gxy[*n*] – дискретная квадратичная функция когерентности
G[*n, k*] – частотный коэффициент передачи, целевая функция фильтра

широкополосного шума для бина n на кадре k

$G(f, k)$ – частотный коэффициент передачи, целевая функция фильтра широкополосного шума на частоте f на кадре k

g – коэффициент передачи

$g[i]$ – коэффициент передачи для дискретного момента времени i

$h[i]$ – функция импульсного отклика

$H(f)$ – амплитудно-частотная характеристика

$H[n]$ – дискретная амплитудно-частотная характеристика

$H\{\ \}$ – преобразование, выполняемое системой

$HR\{x\}$ – оператор полуволнового выпрямления (Half-wave Rectification, функция Хевисайда): $HR\{x\} = ½\ (x + |x|)$

i – дискретное время (временной индекс)

$I(f)$ – интенсивность (сила) звука

$Ix(n) = I(n, k) = |X(n, k)|^2$ – периодограмма сигнала $x[i]$ на кадре k

Im{ } оператор вычисления мнимой части числа, заключенного в скобки

$IDFT\{\ \}$ –дискретное обратное преобразование Фурье

$j = \sqrt{-1}$ – мнимая единица

k – индекс кадров сигнала

L – индекс шага кадров

$L(f)$ – громкость звука

$L\{\ \}$ – оператор линейного преобразования

Lg, lg – десятичный логарифм

Ln, ln – натуральный логарифм

$LPCCx[m]$ – коэффициенты кепстра линейного предсказания

Mx – средневыпрямленное значение сигнала (магнитуда)

$Mx[k]$ – средневыпрямленное значение сигнала на k-м кадре: $Mx[k] = \frac{1}{N}\sum_{i=1}^{N} |x[i]|$

$Mx[i]$ – средневыпрямленное значение сигнала для дискретного момента времени i

$MFCCx(m)$ – мел-спектральные коэффициенты кепстра

$MTF(Fm)$ – модуляционная передаточная функция

m – индекс

N – размерность (количество бинов) дискретного спектра

n – индекс для дискретных частот (бинов) спектра

p – параметр

$Pxx(f)$ – спектральная плотность мощности (спектр мощности)

$Pnn(f)$ – обычно спектр мощности шума

$Pss(f)$ – обычно спектр мощности речевого сигнала

$Pxx[n]$ – дискретная спектральная плотность мощности

$Pxy(f)$ – кросс-спектр (взаимная спектральная плотность мощности)

$Px[i]$ – мощность сигнала для дискретного момента времени i

Px – мощность сигнала

Px[*k*] –мощность сигнала на *k*-м кадре
Ps – мощность речевого сигнала
Pn – мощность шума
PCCxy – коэффициент корреляции Пирсона

q – индекс

Rxx[m] – дискретная автокорреляционная функция, соответствующая временному сдвигу *m*: *Cxx*[*m*]/ *Cxx*[0] (нормированная автоковариация)
Rxy[*m*] – дискретная кросс-корреляционная функция двух процессов (нормированная кросс-ковариация), соответствующая временному сдвигу *m*
Rxx – автокорреляционная матрица, состоящая из элементов Rxx[*i*, *j*]
RXY – коэффициент корреляции между векторами
Re{ } – оператор вычисления реальной части числа, заключенного в скобки

s[*i*] – временной ряд (обычно речевой сигнал)
SNR – отношение сигнал-шум
SNRseg – сегментное отношение сигнал-шум
Sxx(*f*) – спектральная плотность энергии
Sxx(*n*) – дискретная спектральная плотность энергии

t – непрерывное время
T – шаг (интервал, период) дискретизации (выборки) по времени (1/*Fs*)
THR – порог
To – период основного тона, период процесса
Ta – постоянная времени сглаживания

v[*i*] – возбуждающий случайный процесс (в т.ч. белый шум)

W(*f*) – АЧХ фильтра (обычно фильтра Винера)
w[*i*] – Коэффициенты фильтра Винера

x[*i*] – временной ряд, дискретный сигнал (обычно вход системы)
x(*t*) – непрерывный сигнал (функция времени)
x(*iT*) – дискретные временные отсчеты сигнала *x*(*t*)
X[*k*] = [*x*[kL], *x*[kL-1], … *x*[kL-N+1]]T – вектор отсчетов сигнала на кадре
X[*n, k*], Xn[*k*] – коэффициенты дискретного комплексного спектра сигнала на кадре *k*
Xre[*n, k*], Xim[*n, k*] – реальная и мнимая части дискретного комплексного спектра на кадре *k*
|*Xn*[*k*]|, |*X*[*n, k*]| – модуль дискретного комплексного спектра на кадре *k*
<*x*> – оценка средней величины сигнала *x*[*i*]

y[*i*] – временной ряд, дискретный сигнал (обычно выход системы)

z – оператор сдвига вперед: zx[*i*] = *x*[*i*+1]
z-1 – оператор сдвига назад: z-1x[*i*] = *x*[*i*-1]
Zk – частота пересечения нуля сигнала на *k*-м кадре

α – коэффициент забывания /усреднения (*forgetting/averaging factor*) ФЭС
β – постоянная сглаживания (*smoothing factor*) ФЭС
λ – длина звуковой волны
$\delta(t)$ – дельта-функция Дирака
δ[*i*] – единичная цифровая дельта-функция (функция Кронекера)
$\Delta\dot{x}$ – смещение оценки величины *x*
Δ – оператор разности назад первого порядка: Δ*x*[*i*] = *x*[*i*] - *x*[*i*-1] = (1- *z*-1) *x*[*i*]
Δ*F*(BW) – ширина полосы пропускания фильтра / дискретного спектра, частотное разрешение
Δ*T* – временной интервал, длительность окна анализа
ε*t* – возбуждающий случайный процесс (в т.ч. белый шум)
Γ*xy*(*f*), Γ*xy*(*n*) – комплексная функция когерентности
μ*x* – среднее значение случайной величины (параметра) *x*
μ*x*[*k*] – среднее значение случайной величины (параметра) *x* на кадре *k*
ρ*xy* – корреляция (нормированная ковариация) = σ*xy*/ σ*x* σ*y*
σ – стандартное отклонение
τ – временная задержка (запаздывание)
π = 3,14159265…
φ – фазовый угол
Φ*x*(*f*), Φx[*n*] – фазовый спектр (непрерывный и дискретный)
Φ*xy*(*f*), Φ*xy*[*n*] – фазовый кросс-спектр (непрерывный и дискретный)
Ψ*x* – среднее квадратичное значение случайной величины *x*
ω – угловая частота (рад/с)
Ω – нормализованная угловая частота (рад/отсчет)

( )* – символ комплексного сопряжения
$(*)$ – символ свертки
< > – обозначение операции усреднения по времени
$[\,]^{T}$ – обозначение операции транспонирования вектора
$\prod$ – знак произведения
$\sum$ – знак суммы

Столбов Михаил Борисович
Иванов Владимир Леонидович

Анализ и модели речевых сигналов

Учебное пособие

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н. Ф. Гусарова
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49, лит. А