Анализ и синтез устройств кодирования речевого сигнала на основе антропоморфической обработки и синусоидальных моделей

Соискатель: **Лихачёв Денис Сергеевич**

Научный руководитель: д.т.н., профессор
Петровский Александр Александрович

специальность 05.13.05
«Элементы и устройства вычислительной техники и систем управления»

Задача компрессии речевого сигнала

Тип кодера	Скорость потока данных, кбит/с	Субъективная оценка (MOS)	Сложность реализации
Простейшие кодеры формы сигнала (PCM, ADPCM и др.)	32 – 64	4.5 – 4.9	низкая
Кодеры на основе линейного предсказания (CELP, LD-CELP, VSELP, LPC-10e и др.)	2.4 – 16	2.0 – 4.0	средняя
Гибридные кодеры (гармонические), использующие синусоидальную модель (STC-1, STC-2 и др.)	2.4 – 16	2.0 – 3.5	высокая
Негармонические кодеры речи на основе синусоидальной модели	более 32	**4.5 – 4.8**	средняя

На данный момент проблема качественной компрессии и передачи речи по цифровым каналам со скоростью меньше 6 кбит/с должным образом всё ещё не решена и представляет значительный научный и практический интерес в таких областях, как сотовая связь, передача данных по компьютерным сетям, обработка и хранение речевых сообщений в портативных мультимедиа устройствах

Задача компрессии речевого сигнала

У современных синусоидальных кодеров и кодеров на основе линейного предсказания с увеличением степени компрессии (при скорости потока данных менее 6 кбит/с) качество речи сильно деградирует

С точки зрения качества синтезированного сигнала одними из наиболее перспективных в настоящее время являются **негармонические кодеры речи на основе синусоидальной модели**

Синусоидальная модель речевого сигнала характеризуется:

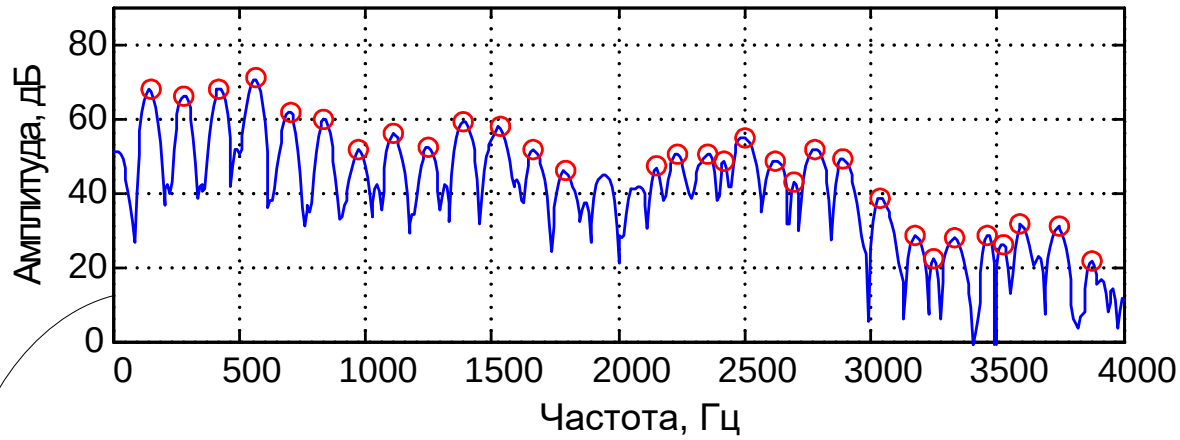
1) **высоким качеством** синтезированной речи,
2) возможностью применения для кодирования широкополосных сигналов
3) возможностью изменения характеристик речи в процессе синтеза

Недостатки: высокая вычислительная сложность и большое количество параметров

Кодирование сигнала на основе синусоидальной модели

McAulay R.J., Quatieri T.F.

$$s(n) = \sum_{l=1}^{L} A_l \sin(\omega_l n + \varphi_l)$$



Входной сигнал

s(n)

Разложение речевого сигнала на синусоидальные составляющие

Определение синусоидальных параметров

Амплитуды

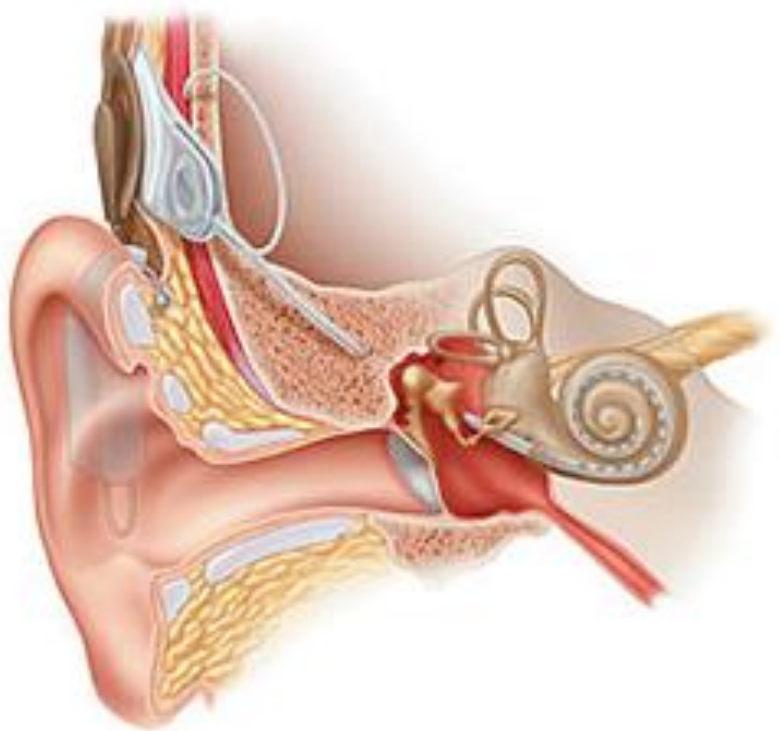
Частоты

Фазы

Экспериментальные результаты показывают, что для представления речи с высоким качеством необходимо 60-80 синусоид на 20 мс отрезок речевого сигнала, что в сумме даёт **200 параметров / фрейм**.

Кохлеарная имплантация

Кохлеарная имплантация – в улитку уха человека вводится набор электродов, через который проводится прямая стимуляция слухового нерва электрическими импульсами, определенным образом соответствующими воспринимаемому акустическому сигналу.



Количество электродов

В настоящее время считается, что для практически безошибочного распознавания акустической информации и идентификации диктора достаточно:

8 электродов – для речевого сигнала,

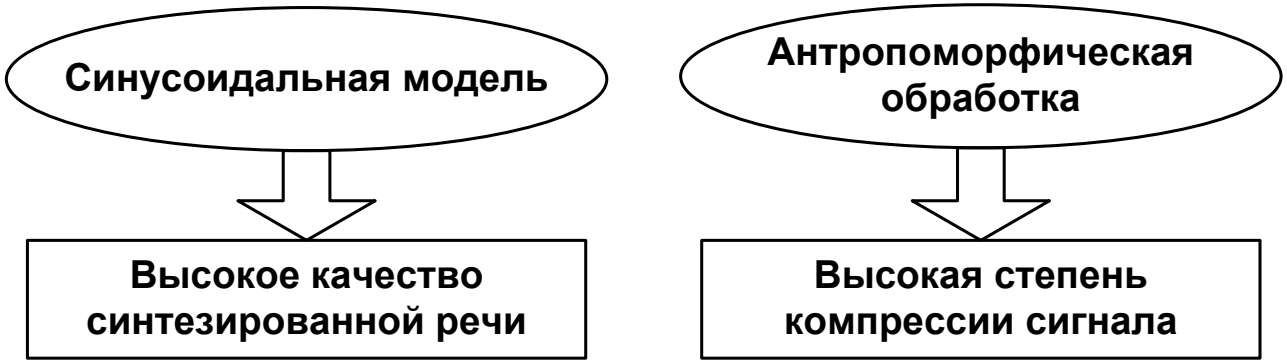
20-22 – для широкополосного.

Анализ и синтез речевого процессора улитковой имплантации: дис. ... канд. техн. наук: 05.13.05 / Ярослав Башун (науч. рук.: д.т.н., проф. Петровский А.А.). – Минск, 2001.

Постановка задачи

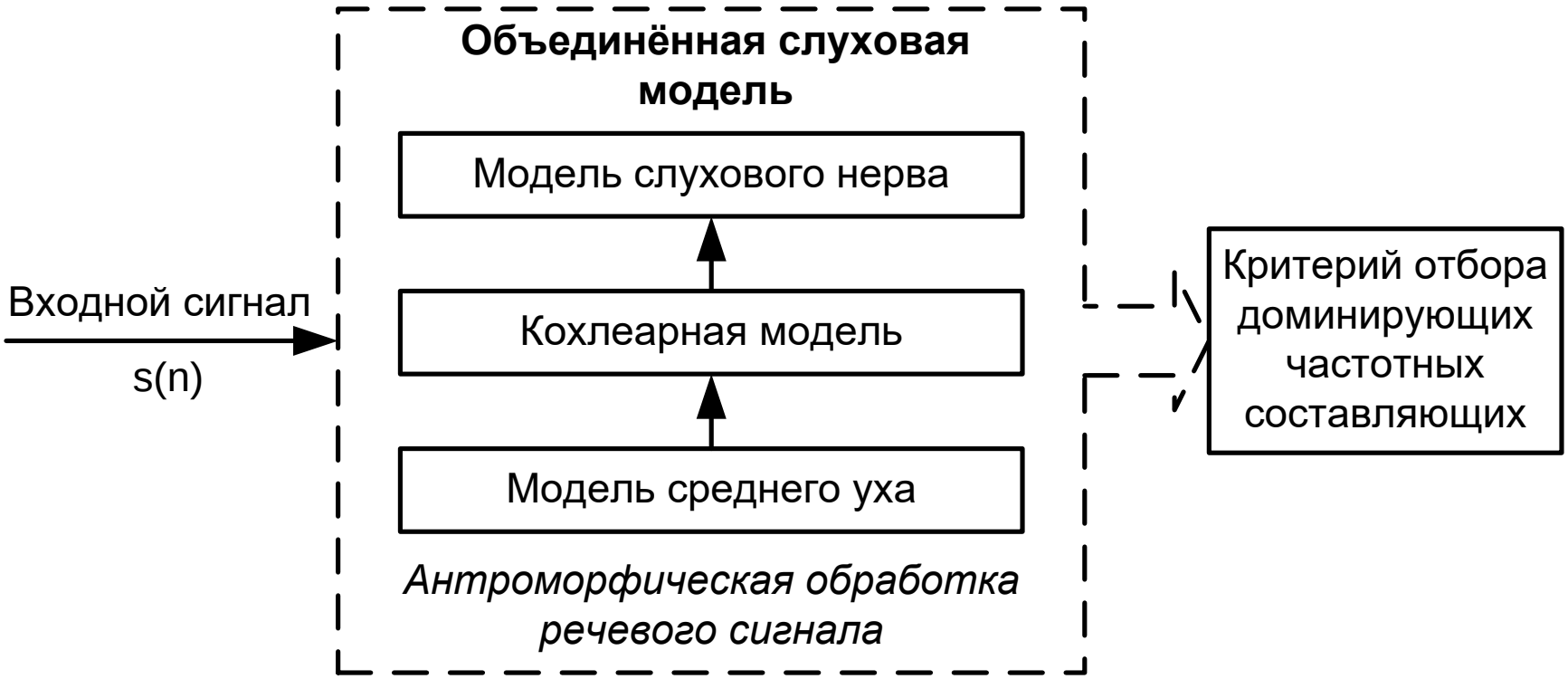
Необходимо получить ***высокую степень компрессии*** при хорошем ***качестве восстановленной речи***

Гипотеза: предполагается, что использование антропоморфической обработки для анализа речевого сигнала позволит выбрать только такие частотные компоненты, которых будет достаточно для качественного кодирования речи на основе синусоидальной модели.



Целью работы является разработка методов и алгоритмов анализа и синтеза устройств кодирования речи с высокой степенью компрессии, использующих антропоморфическую обработку речевого сигнала.

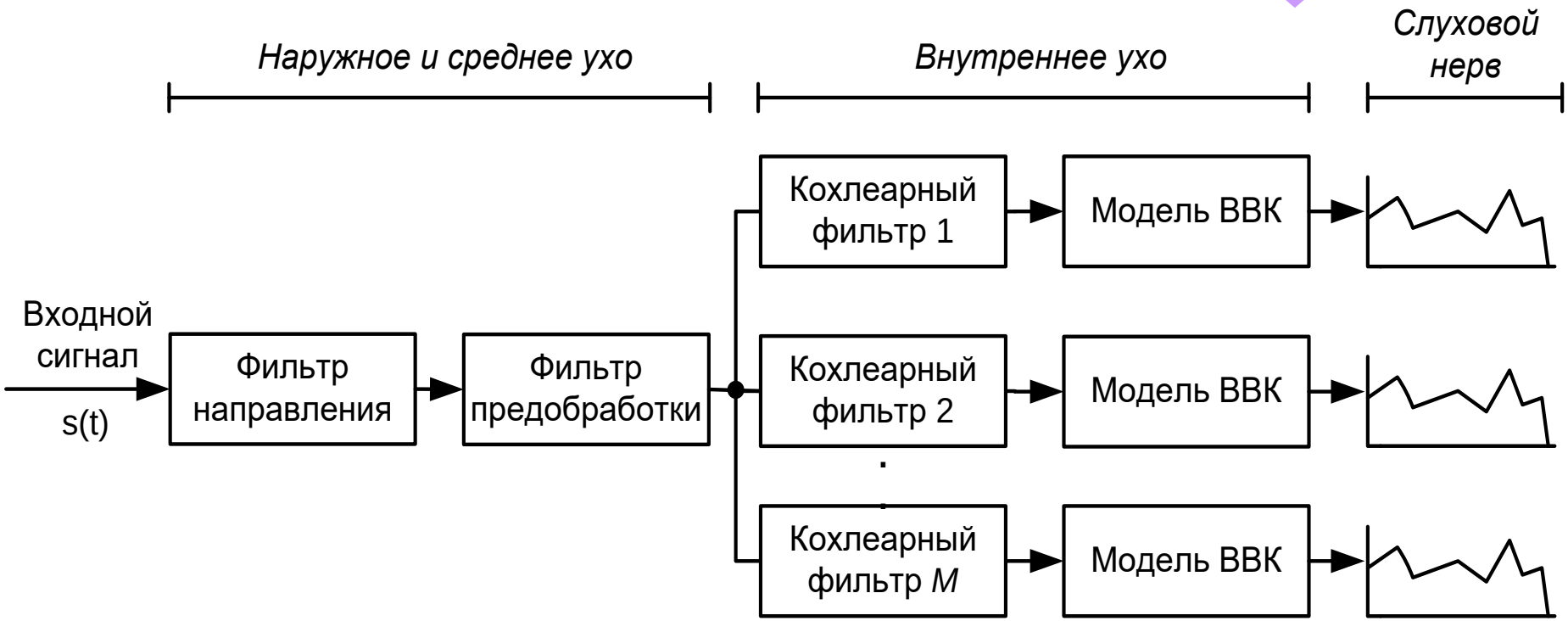
Задача антропоморфической обработки речевого сигнала



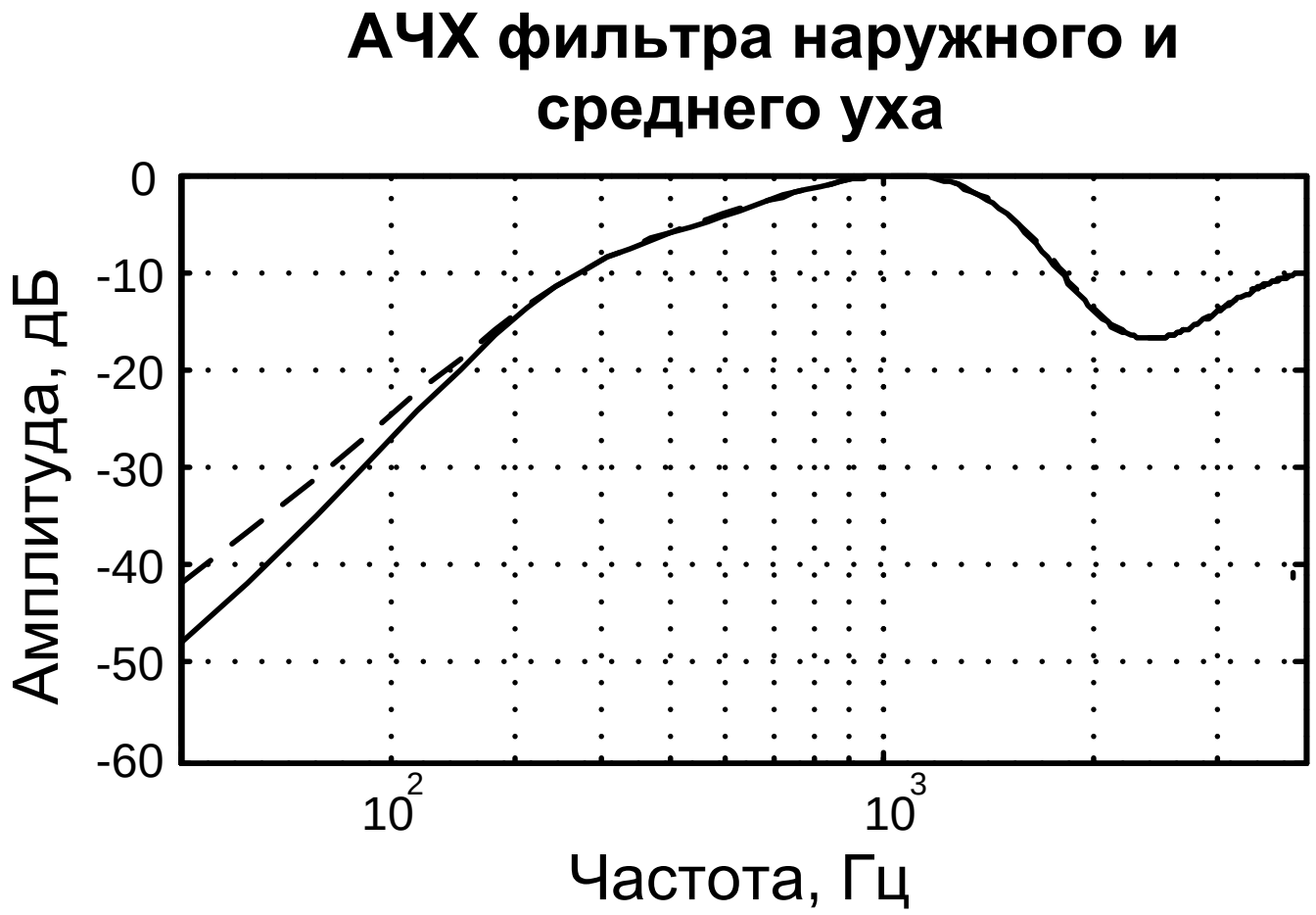
Речевой сигнал представляется ограниченным количеством **“доминирующих”** частотных компонент, а для выработки критерия их отбора используется **антропоморфическая обработка** на основе трёх слуховых моделей: **модели внешнего и среднего уха**, **кохлеарной модели** (пассивной и активной) и **модели слухового нерва человека**

Структурная схема слуховой модели

На основании анализа существующих методов антропоморфической обработки цифровых сигналов показывается, что структура слуховой модели с высокой степенью адекватности реальным физиологическим процессам имеет следующий вид



Модель внешнего и среднего уха



сплошная линия – экспериментальные данные Гласберга и Мура (B.R. Glasberg, B.C.J. Moore)
штриховая линия – аппроксимация с помощью БИХ-фильтра 20-го порядка

Пассивная кохлеарная модель

- В качестве основной модели периферической части слуховой системы человека используется SDCM-модель – Second Order Difference Cochlea Model (разностная кохлеарная модель второго порядка):

$$y_m(n)+b_{1m}y_m(n-1)+b_{2m}y_m(n-2)=A_m a_{0m}[u_s(n)-u_s(n-2)]$$

где $y_m(n)$ – смещение базилярной мембраны в позиции x_m ;
$u_s(n)$ – входной синусоидальный сигнал, характеризующий скорость перемещения стремечка

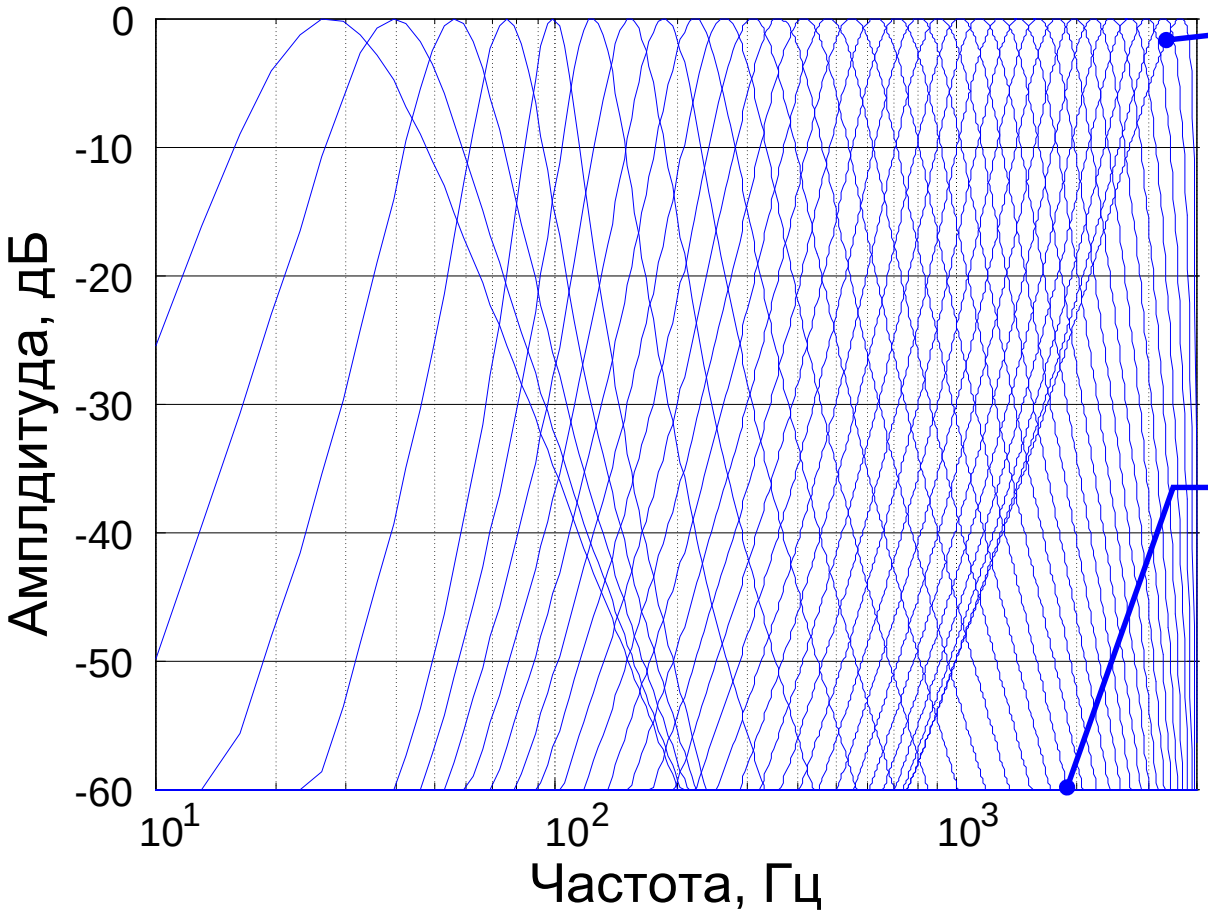
Передаточная функция модели улитки в дискретном пространстве и времени

$$H_m(z)=\left[A_m\frac{a_{0m}\left(1-z^{-2}\right)}{1+b_{1m}z^{-1}+b_{2m}z^{-2}}\right]^{NS}$$

b_{1k} a_{0k}
b_{2k} A_k

Данные параметры определяются физическими свойствами базилярной мембраны в данном месте и изменяются вдоль базилярной мембраны

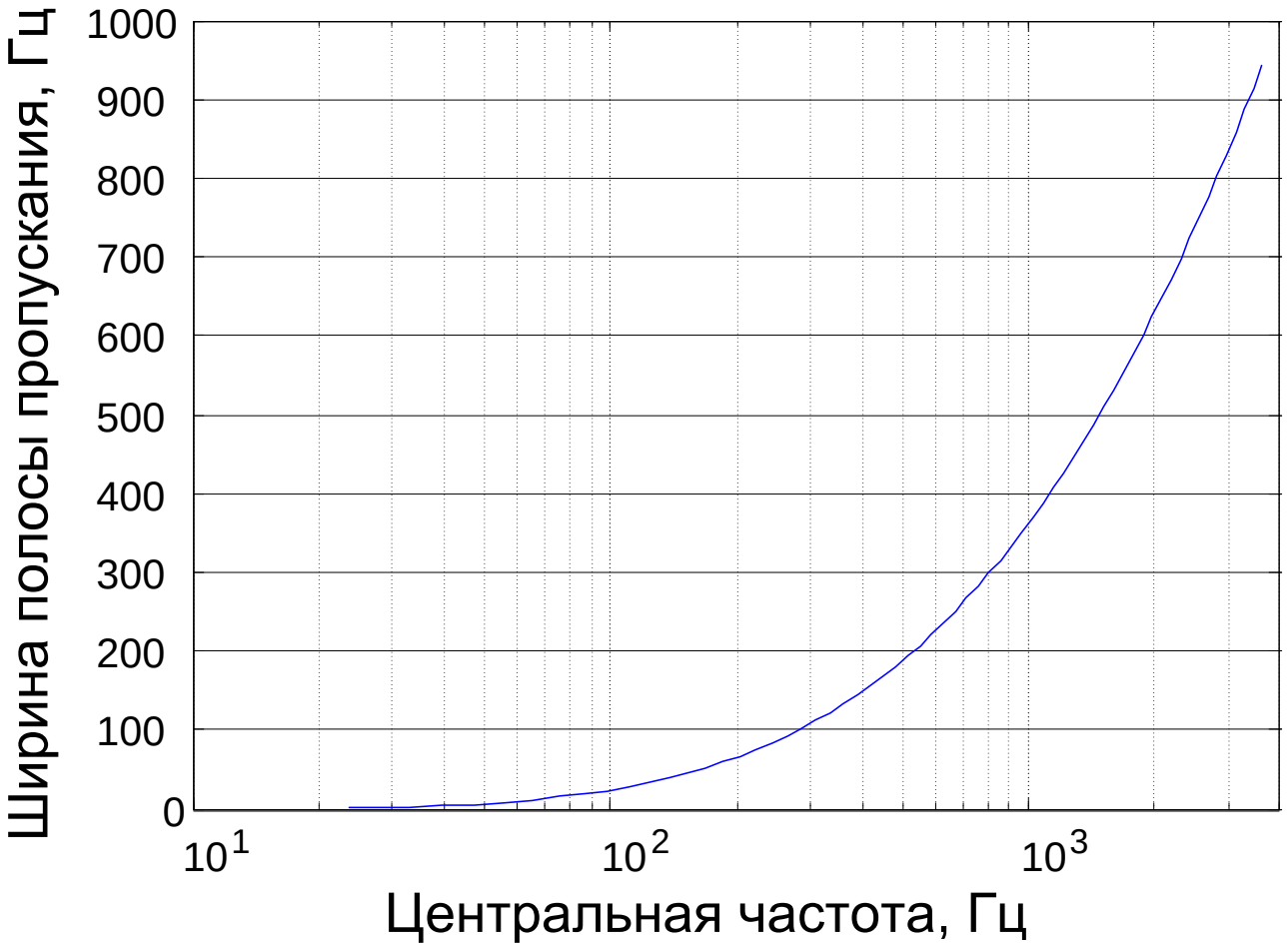
Амплитудно-частотные характеристики для 32 кохлеарных фильтров



Частотная характеристика каждого фильтра отражает механические свойства базилярной мембраны в данном месте

Центральные частоты фильтров определяются соответствующими *характеристическими частотами*

Характеристика полосы пропускания по уровню 3 дБ

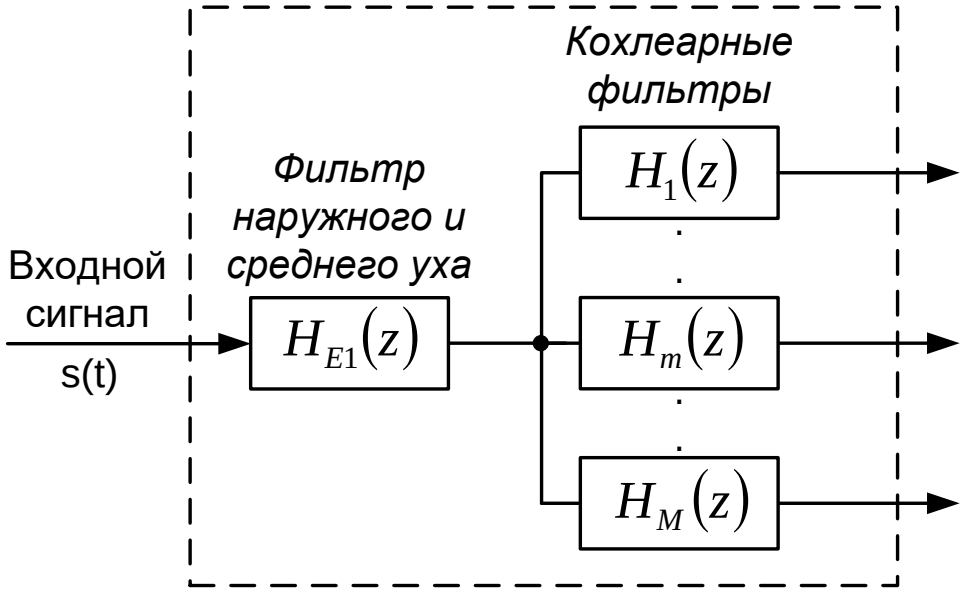


- С увеличением центральной частоты увеличивается и полоса пропускания фильтров

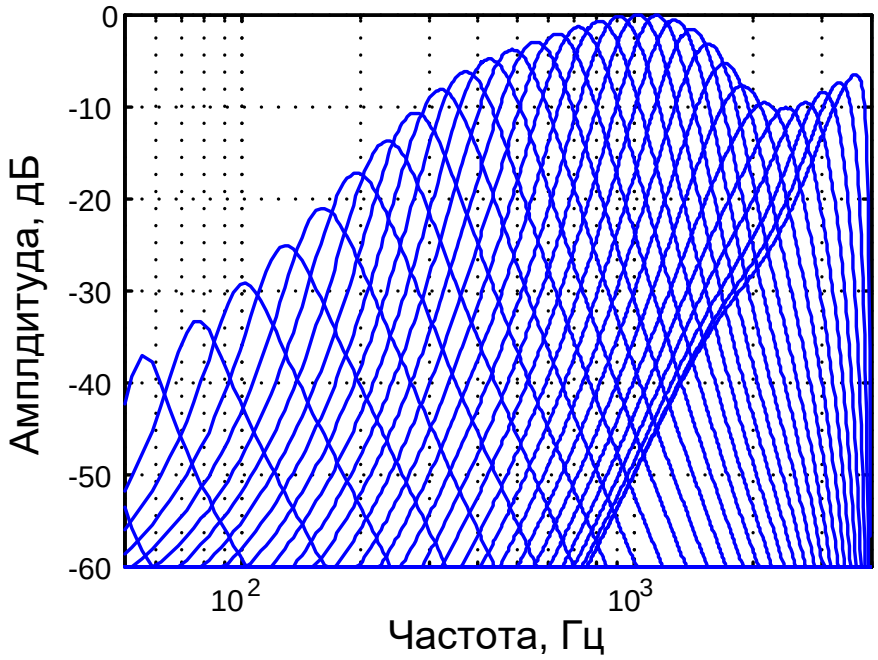
Модифицированная кохлеарная модель

С целью повышения степени адекватности реальным физиологическим процессам предлагается провести коррекцию кохлеарной модели с учётом передаточной характеристики внешнего и среднего уха

Схема включения фильтра наружного и среднего уха и кохлеарных фильтров



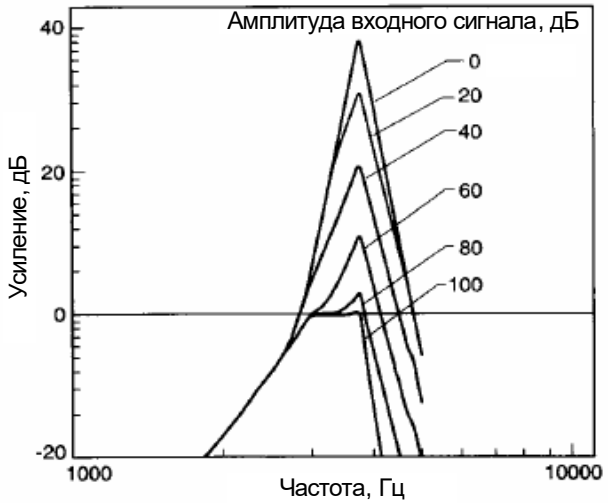
АЧХ для 32 модифицированных кохлеарных фильтров



Для повышения точности кохлеарного анализа предложено использовать ДПФ с неравномерным частотным разрешением

Активная кохлеарная модель в частотной области

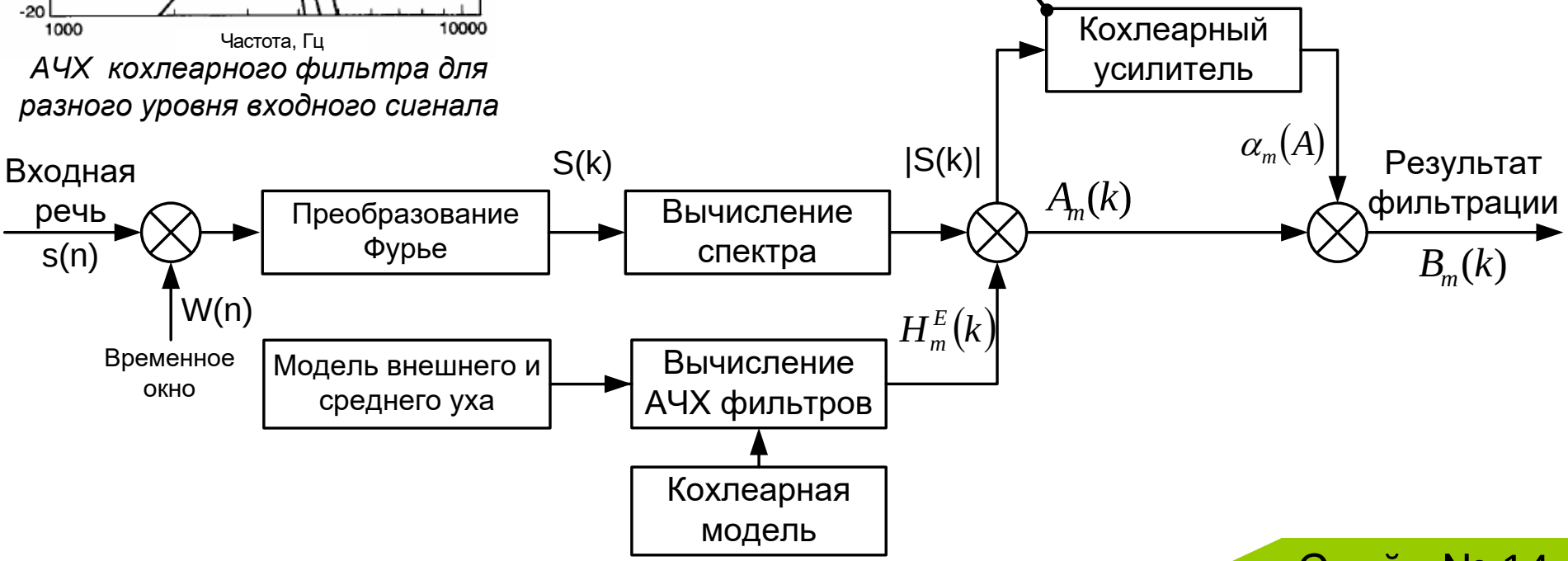
J.L. Goldstein, O. Ghitza



АЧХ кохлеарного фильтра для разного уровня входного сигнала

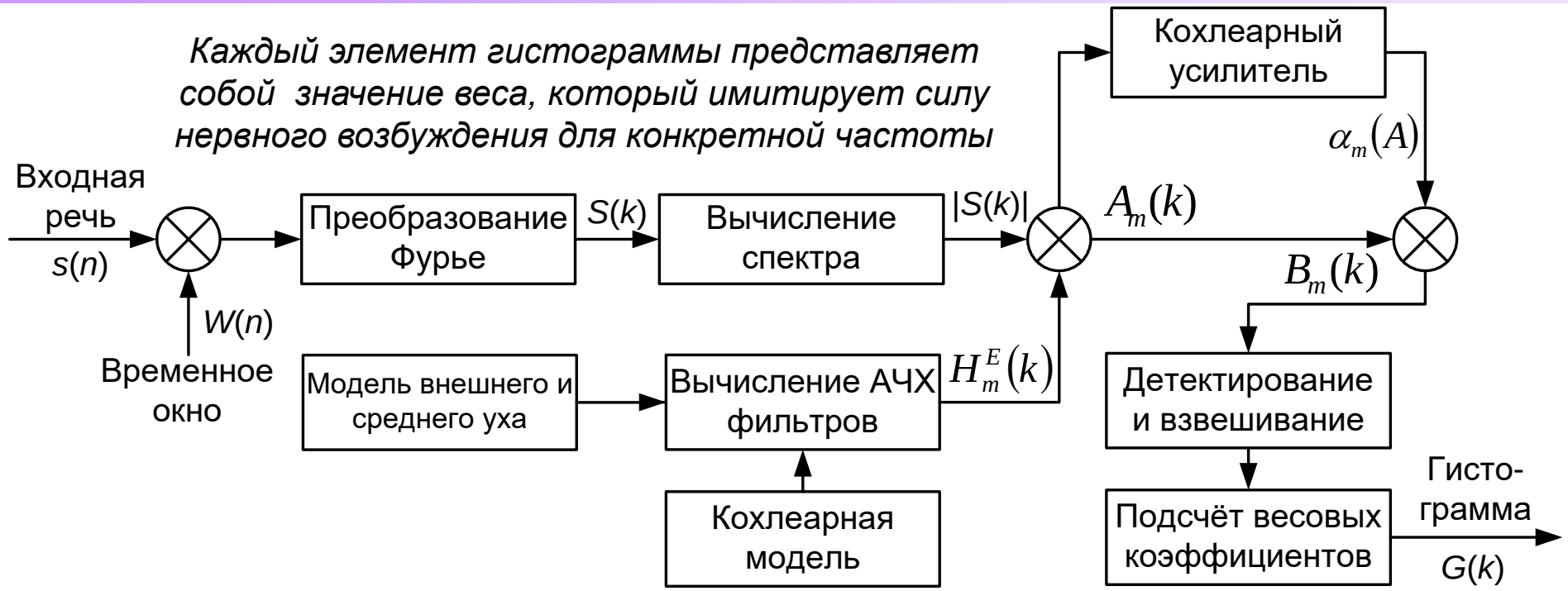
Современные физиологические исследования показывают, что в улитке уха человека происходит некий активный процесс, который приводит к тому, что частотная характеристика кохлеарного фильтра зависит ещё и от интенсивности сигнала на его входе

$$\alpha_m(A) = \frac{100}{10^{40 \cdot \log_{10}\left[|S(k)| \cdot H_m^E(k)\right]}}$$



Модель слухового нерва – слуховая гистограмма

Каждый элемент гистограммы представляет собой значение веса, который имитирует силу нервного возбуждения для конкретной частоты



- Для получения и интерпретации отклика в слуховом нерве предлагается использовать процедуру детектирования и взвешивания, с помощью которой описывается поведение нервных волокон, раздражаемых во внутренних волосковых клетках
- Частотные составляющие анализируемого речевого сигнала предлагается дифференцировать по степени их важности для человеческого слуха с помощью **слуховой гистограммы**

Вычисление слуховой гистограммы (1)

Гистограмма $G(k)$ вычисляется с помощью следующих выражений:

$$G(k)=\sum_{m=1}^{M_F} G_m(k),$$

$$G_m(k)=|S(k)|\cdot H_m^E(k)\cdot \alpha_m(A)\cdot \sum_{l=1}^{L} p_m(l),$$

$$p_m(l)=\begin{cases}\beta_l, & если\ B_m(k)\geq Amp_l,\\ 0, & если\ B_m(k)<Amp_l,\end{cases}$$

M_F – число кохлеарных фильтров;

$p_m(l)$ – весовая функция;

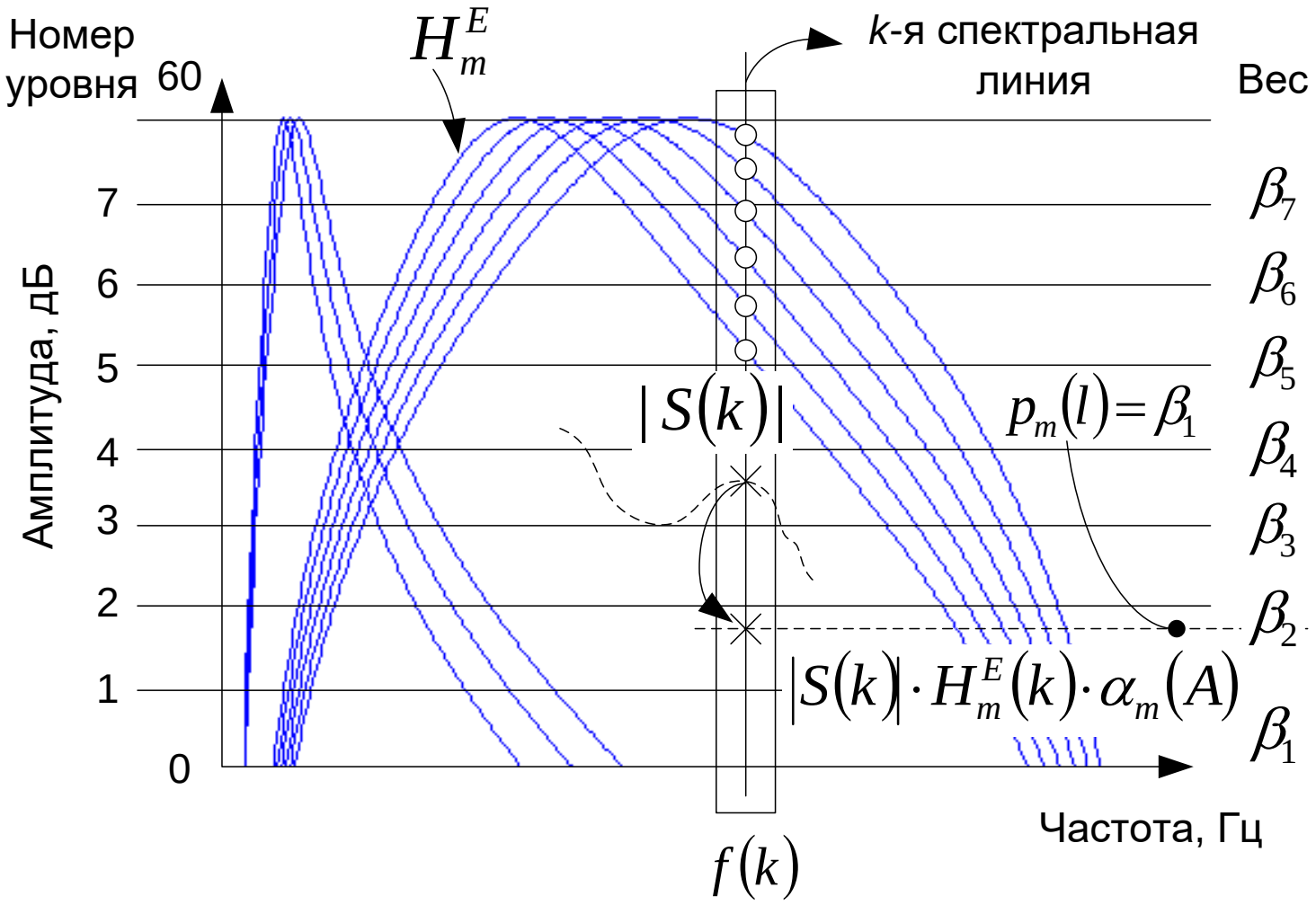
$\alpha_m(A)$ – коэффициент усиления для *m*-го кохлеарного фильтра;

Amp_l – амплитуда *l*-го уровня; L – количество уровней;

β_l – постоянная величина, характеризующая степень нервного возбуждения, каждому уровню сопоставляется своё собственное значение

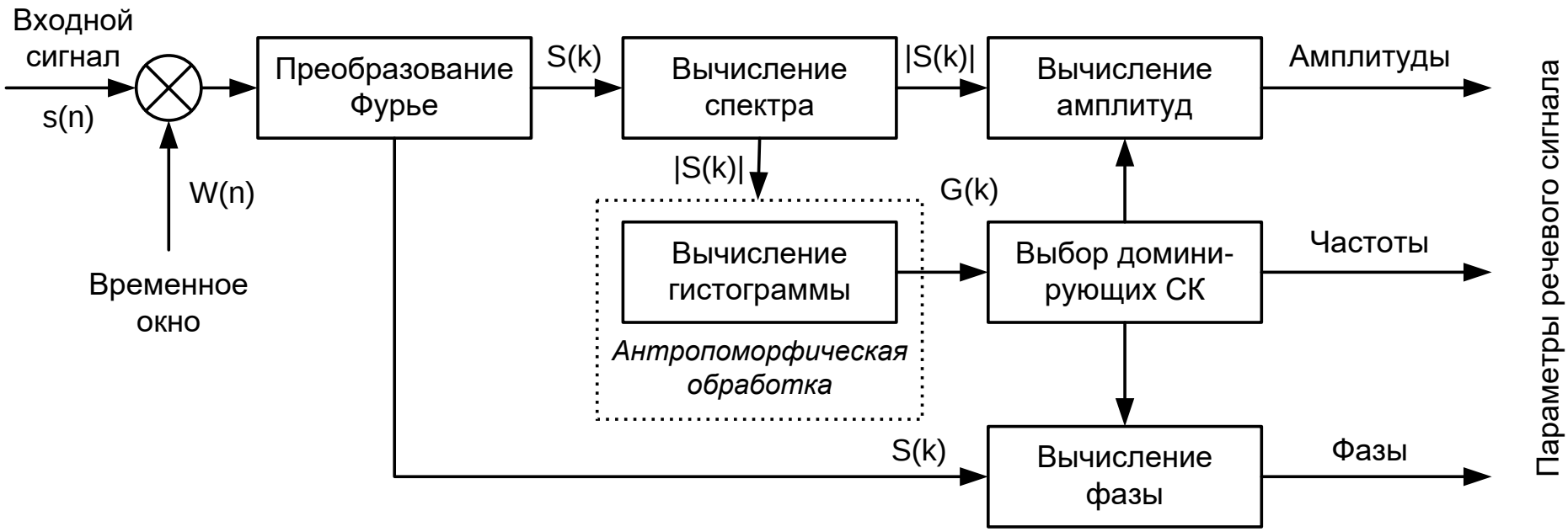
$H_m^E(k)$ – АЧХ модифицированного кохлеарного фильтра

Вычисление слуховой гистограммы (2)



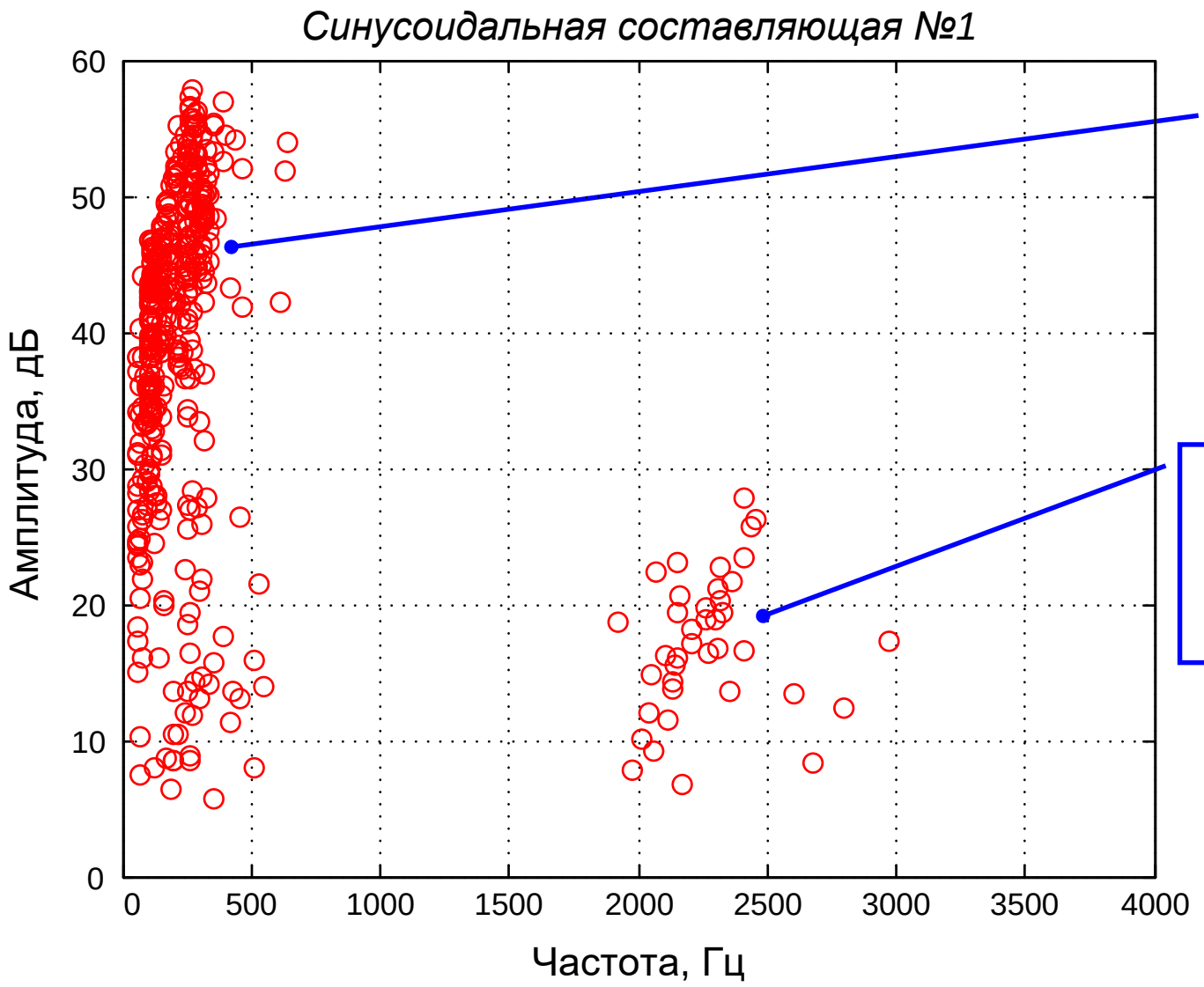
Вес выбирается в зависимости от величины $|S(k)|\cdot H_m^E(k)\cdot\alpha_m(A)$

Устройство анализа речевого сигнала – структура



- **Цель анализа речи** – выделение *доминирующих синусоидальных компонент* (СК) и определение их параметров: амплитуд, частот и фаз, которые затем используются при синтезе
- **Синтез речи** сводится к генерации и суммированию синусоидальных компонент с найденными в процессе анализа параметрами

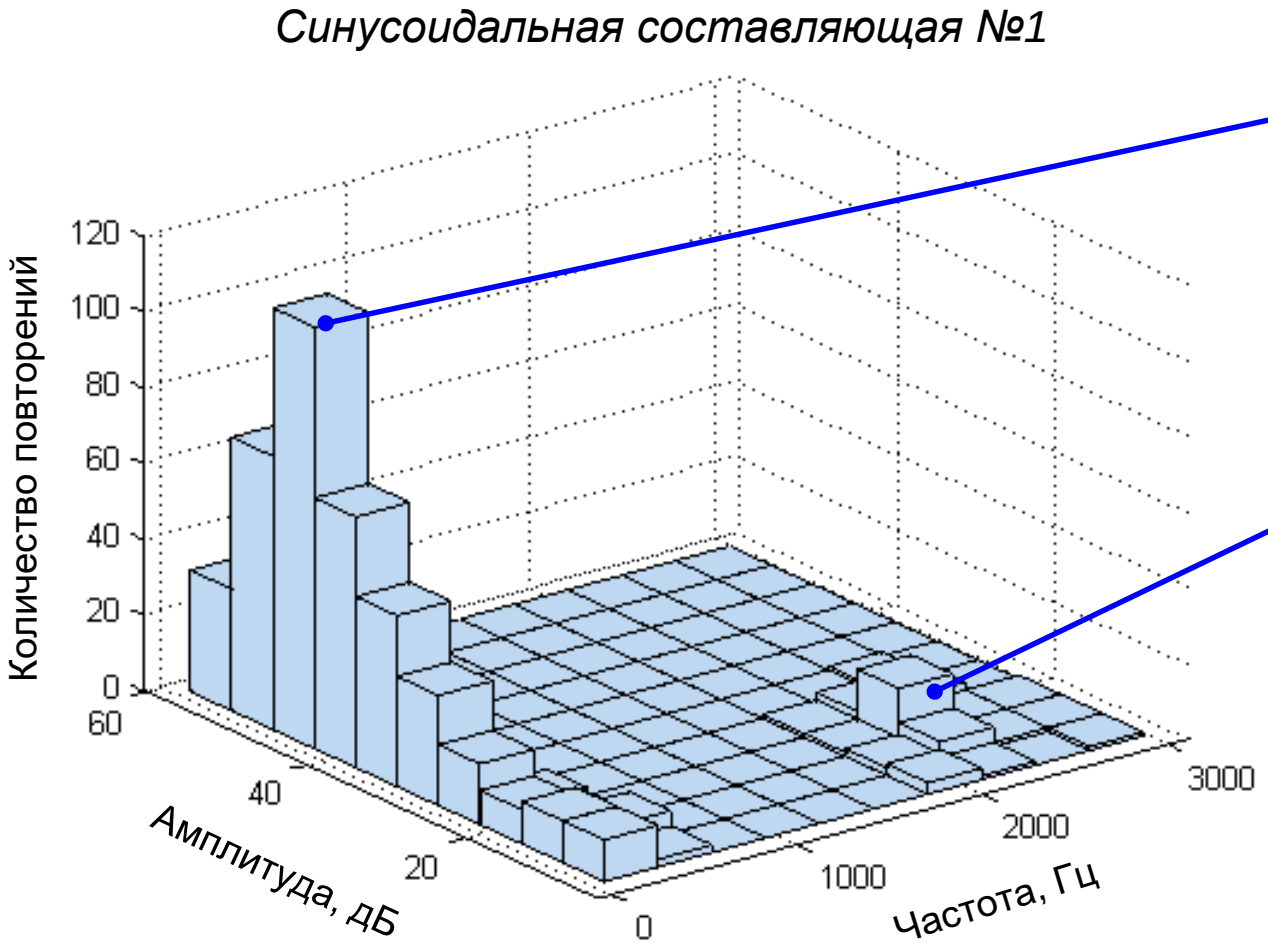
Распределение амплитуд и частот для первой синусоидальной составляющей



Частоты и амплитуды синусоид для вокализованных фреймов

Частоты и амплитуды синусоид для шумовых фреймов

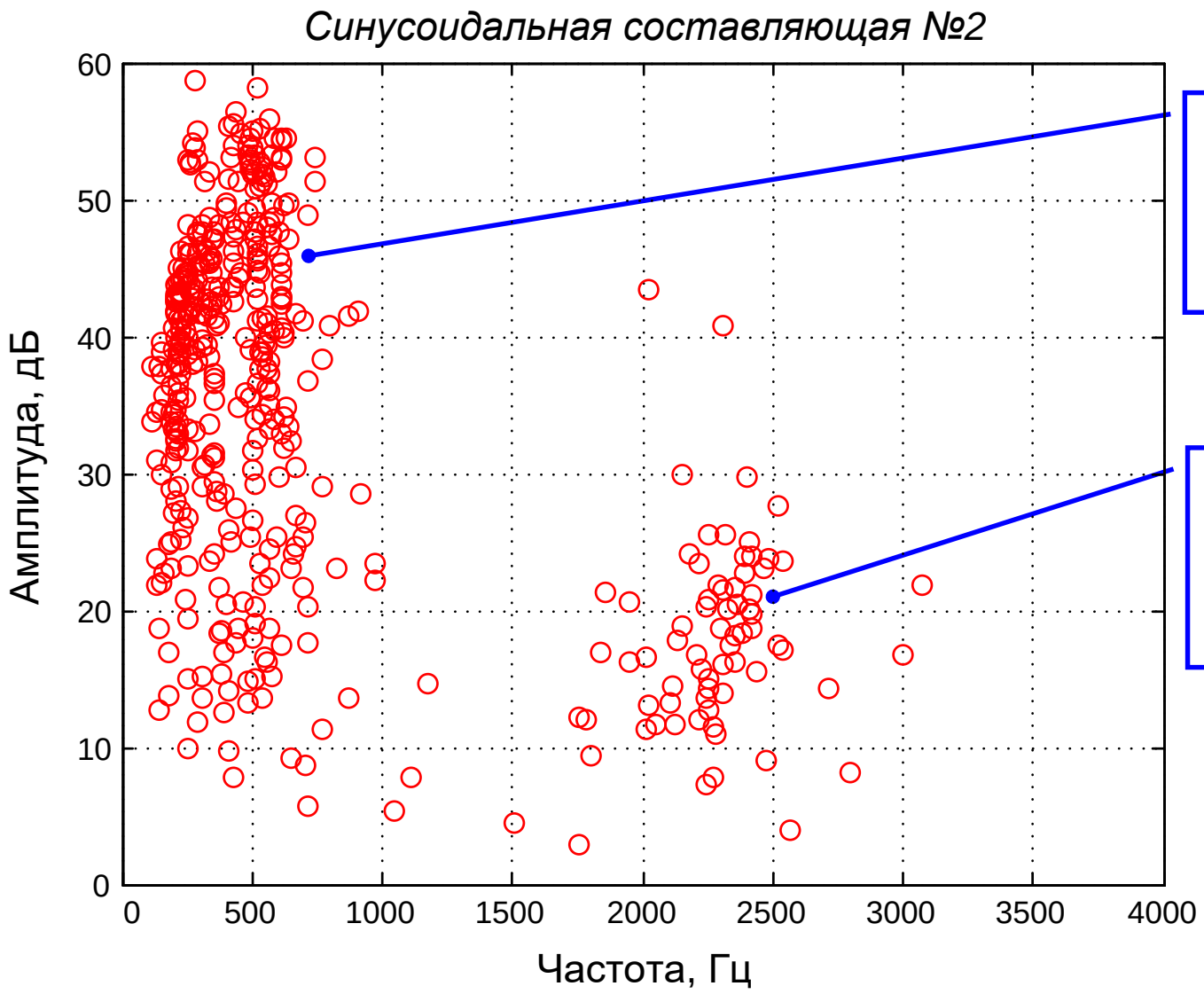
Гистограмма распределения амплитуд и частот для первой синусоидальной составляющей



Распределение параметров локализовано, т.е. максимальные амплитуды в области низких частот

На высоких частотах амплитуды небольшие

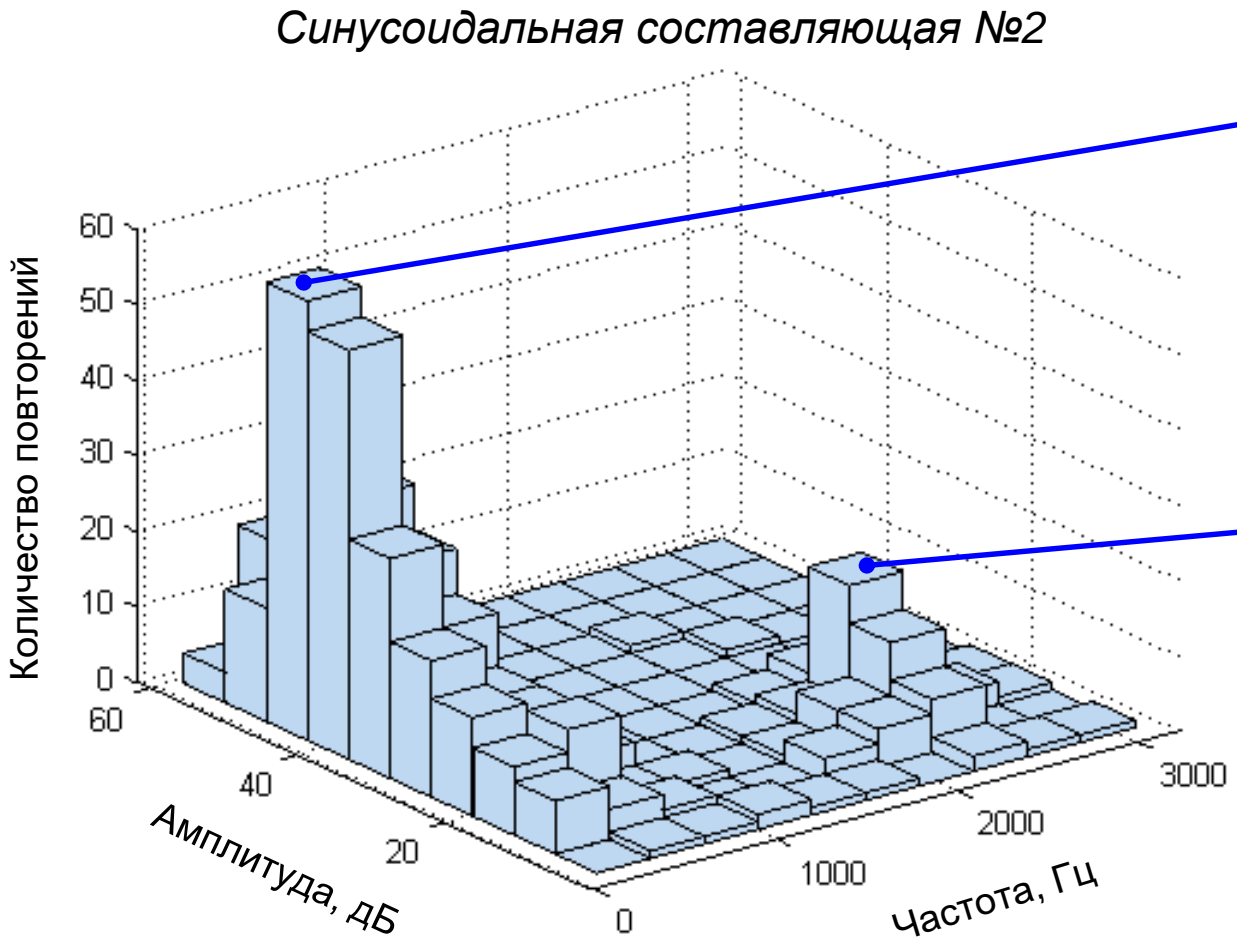
Распределение амплитуд и частот для второй синусоидальной составляющей



Частоты и амплитуды синусоид для вокализованных фреймов

Частоты и амплитуды синусоид для шумовых фреймов

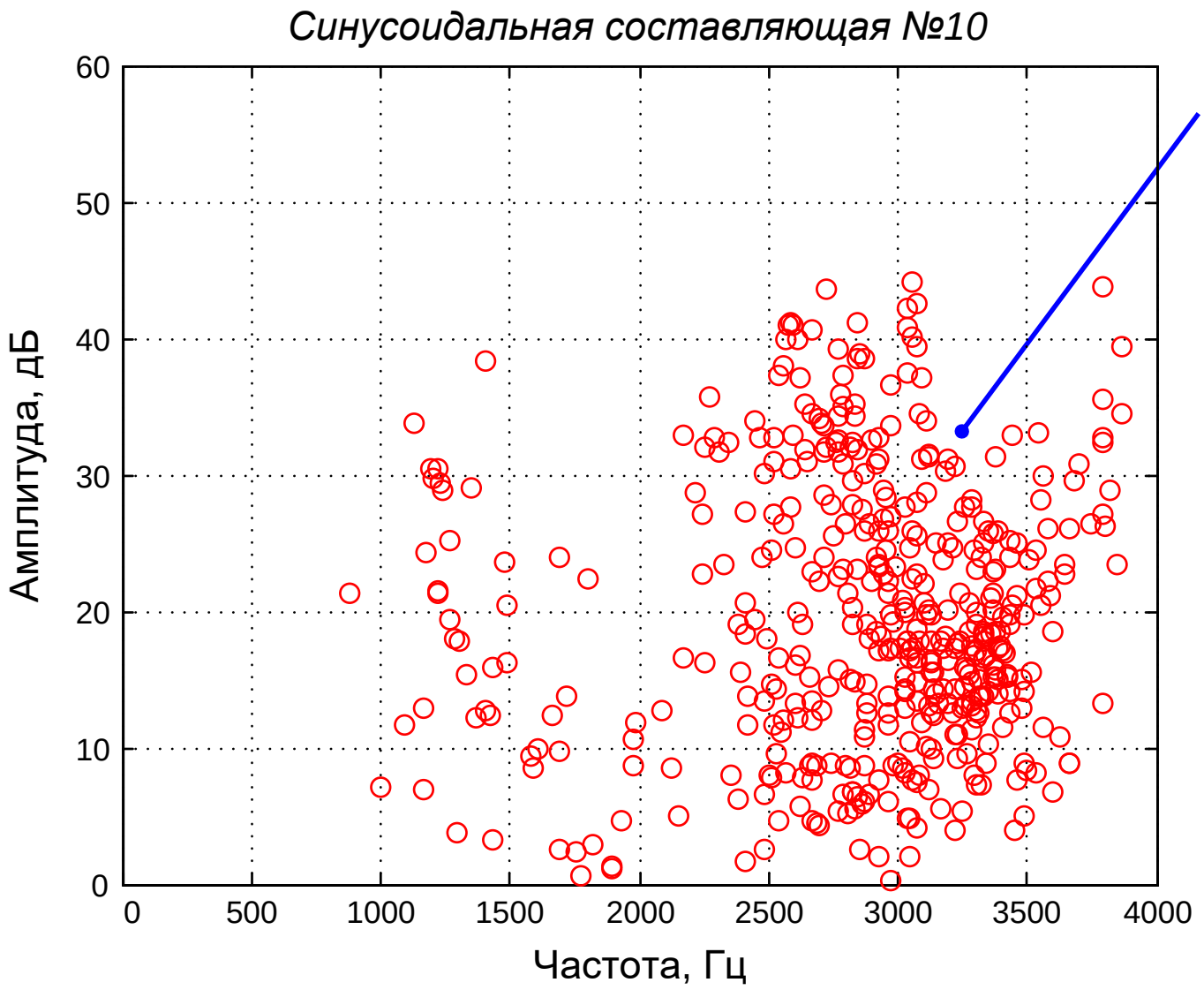
Гистограмма распределения амплитуд и частот для второй синусоидальной составляющей



Область локализации на низких частотах больше, чем для первой составляющей

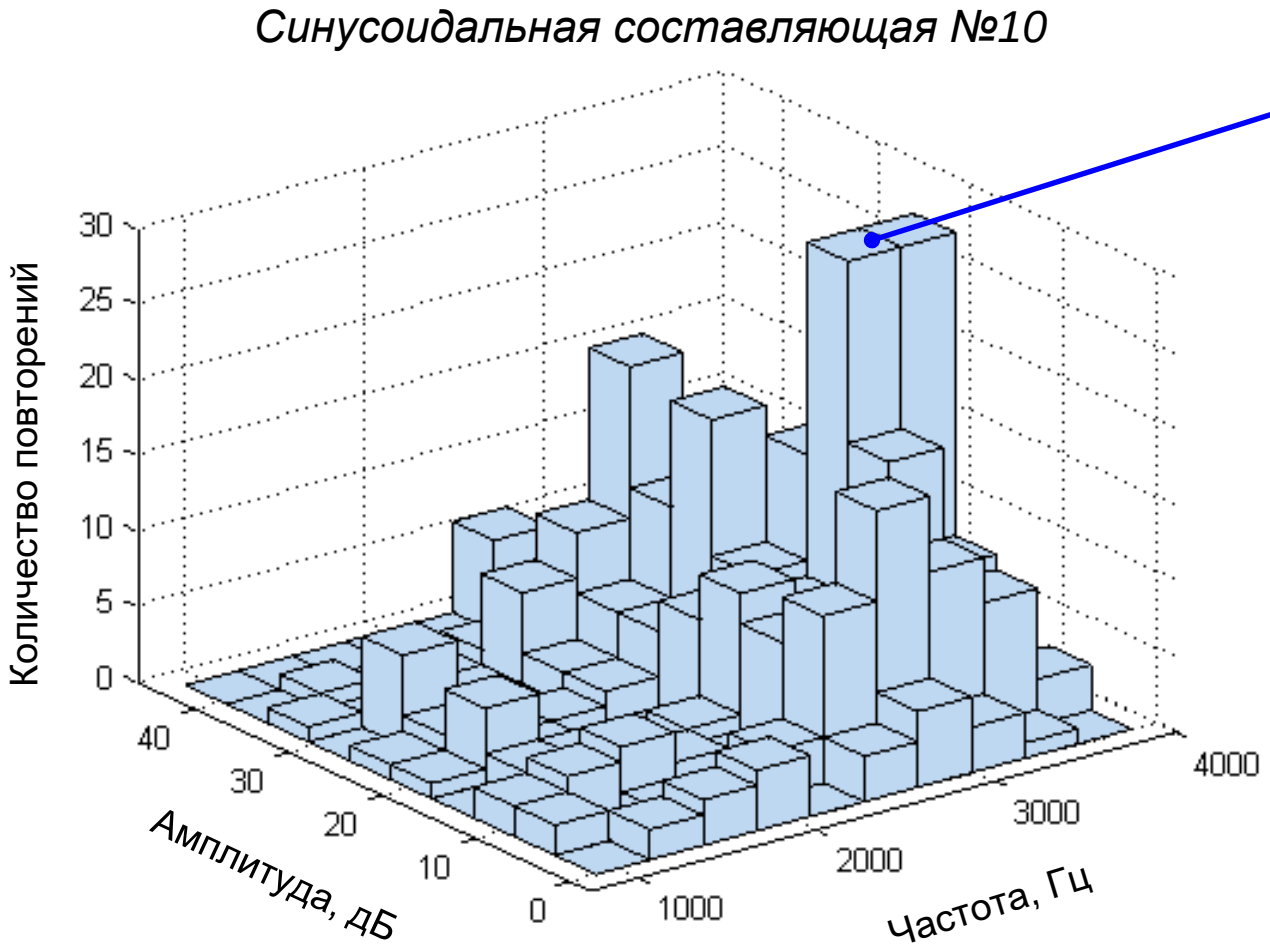
Область локализации на высоких частотах тоже больше, чем для первой составляющей

Распределение амплитуд и частот для десятой синусоидальной составляющей



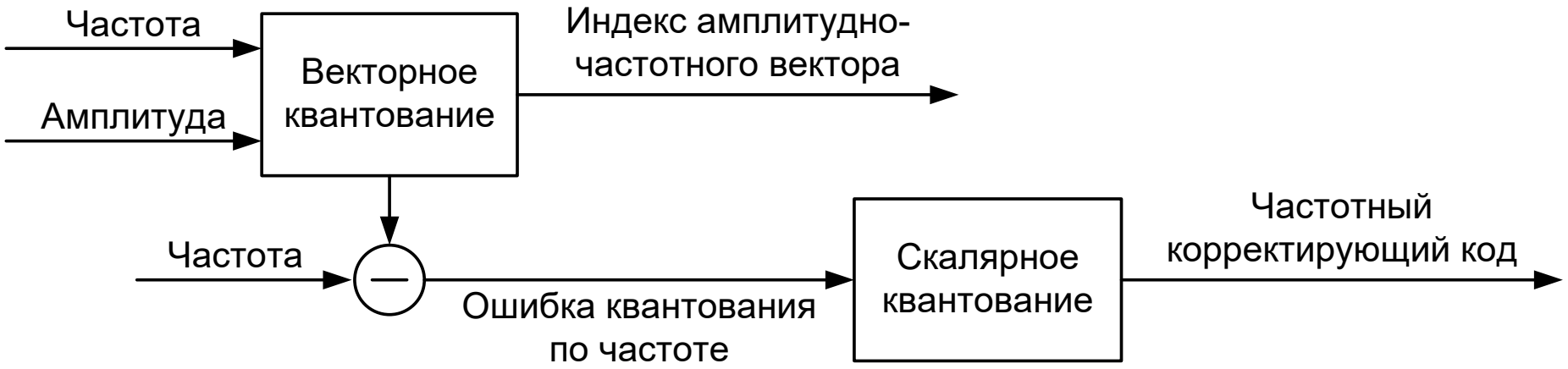
Почти все частоты и амплитуды синусоид сконцентрированы в области высоких частот (2500-3500 Гц)

Гистограмма распределения амплитуд и частот для десятой синусоидальной составляющей



Почти все амплитуды синусоид сконцентрированы в области высоких частот (2500-3500 Гц)

Векторное квантование с частотной коррекцией



- Фазы предлагается кодировать используя **скалярное квантование**, а амплитуды и частоты с использованием **векторного квантования**.

- Экспериментальным образом установлено, что ошибка по частоте чрезвычайно сильно влияет на качество синтезируемой речи, поэтому векторное квантование амплитуд и частот комбинируется со скалярным квантованием ошибки по частоте

Схема квантования параметров синусоидального вокодера с определением соответствующего вклада в поток данных

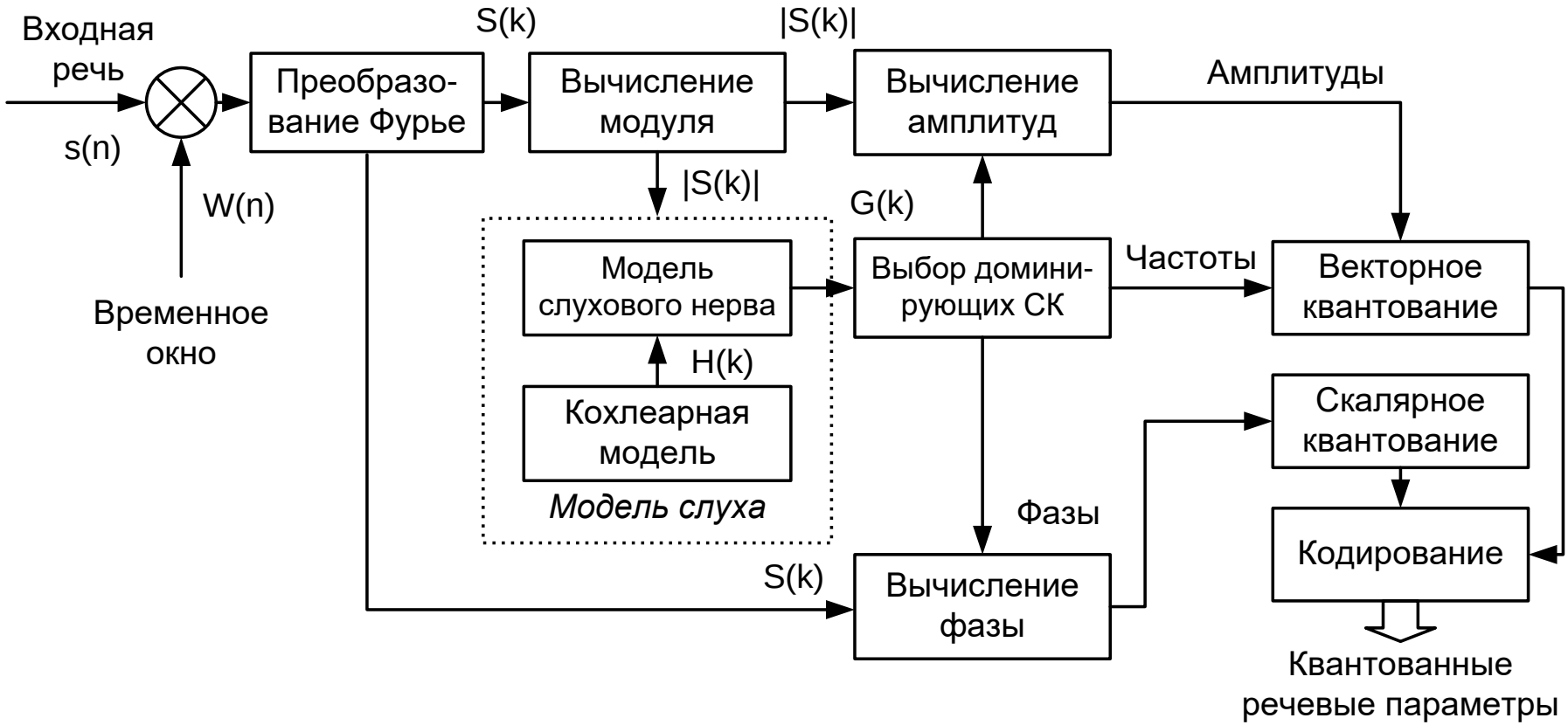
Векторное квантование амплитуд и частот, скалярное квантование фаз

Количество бит для квантование амплитуды и частоты	**Количество бит для квантование фазы**	**Количество бит на одну синусоиду**
13	3	16

Скорость передачи в зависимости от количества доминирующих составляющих

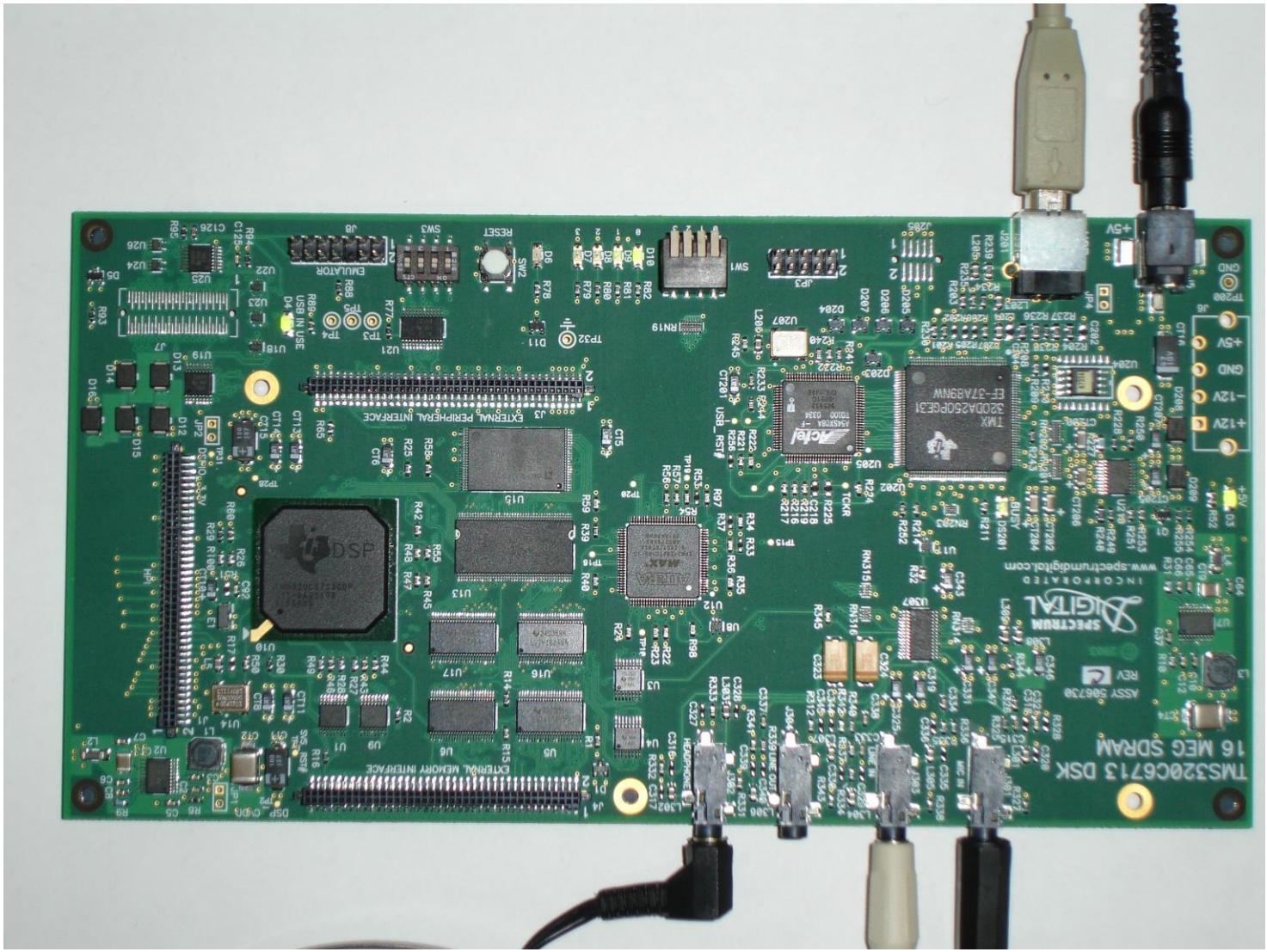
Количество синусоидальных составляющих	**Суммарное количество бит/фрейм**	**Количество фреймов/с**	**Скорость потока, бит/с**
10	160	50	8000
6	96	50	4800

Устройство компрессии с антропоморфической обработкой речевого сигнала: кодер

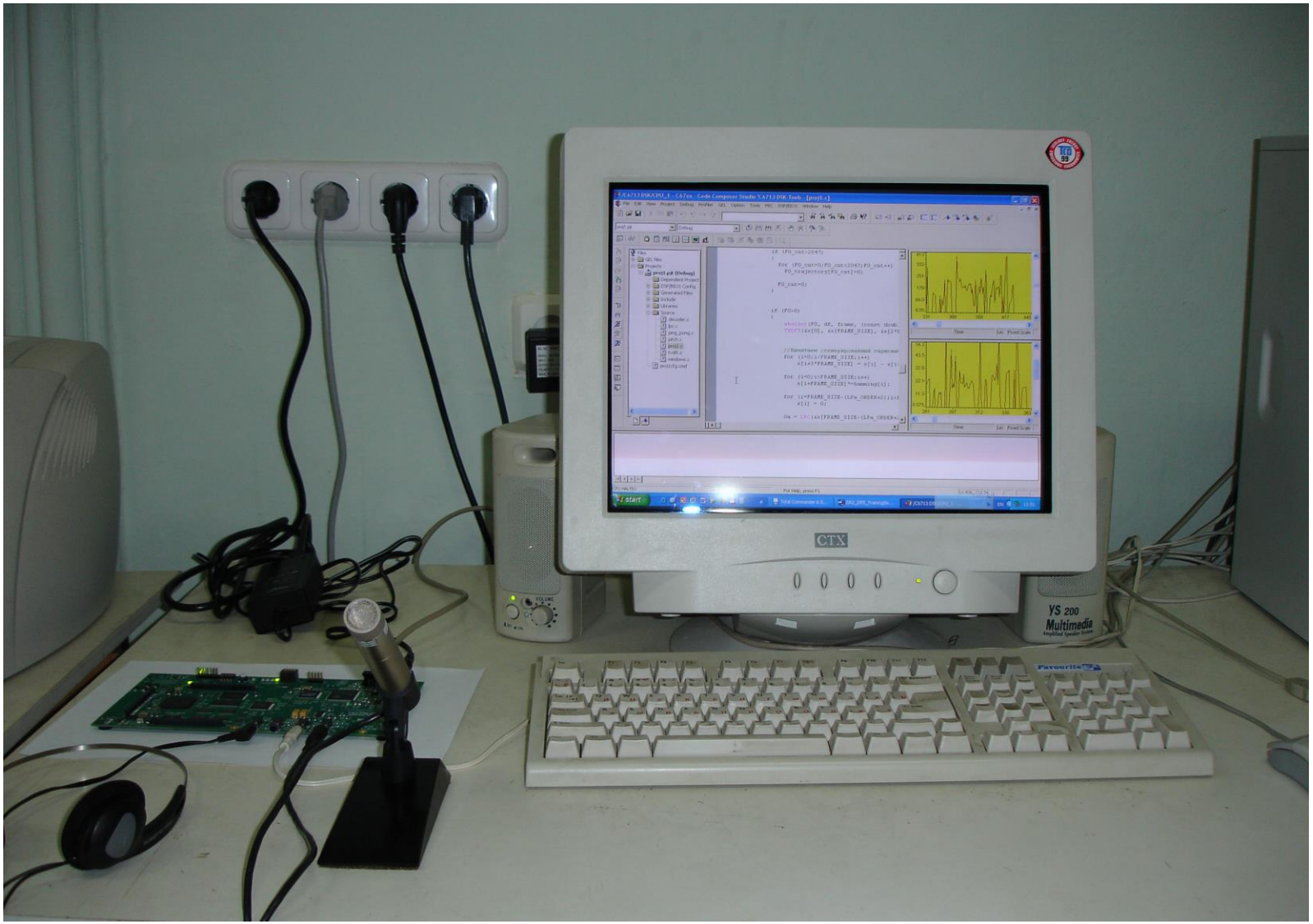


- В качестве базы для аппаратной реализации в реальном масштабе времени используется процессор с VLIW-архитектурой TMS320C6713, позволяющий производить распараллеливание вычислений на уровне команд

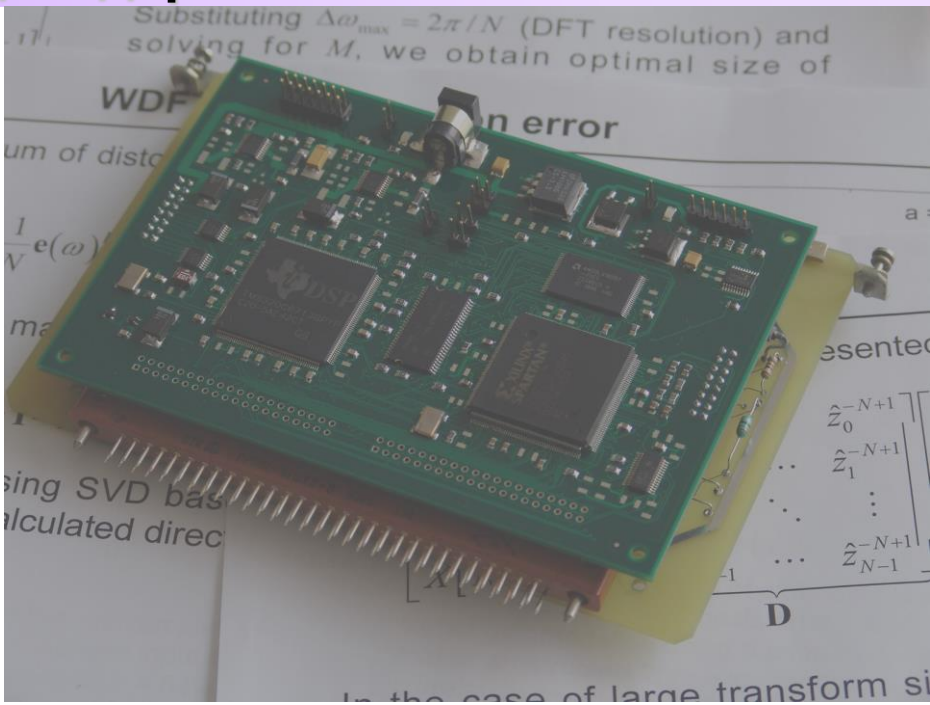
Отладочная плата TMS320C6713 DSK (на базе ЦПОС TMS320C6713 225 МГц)

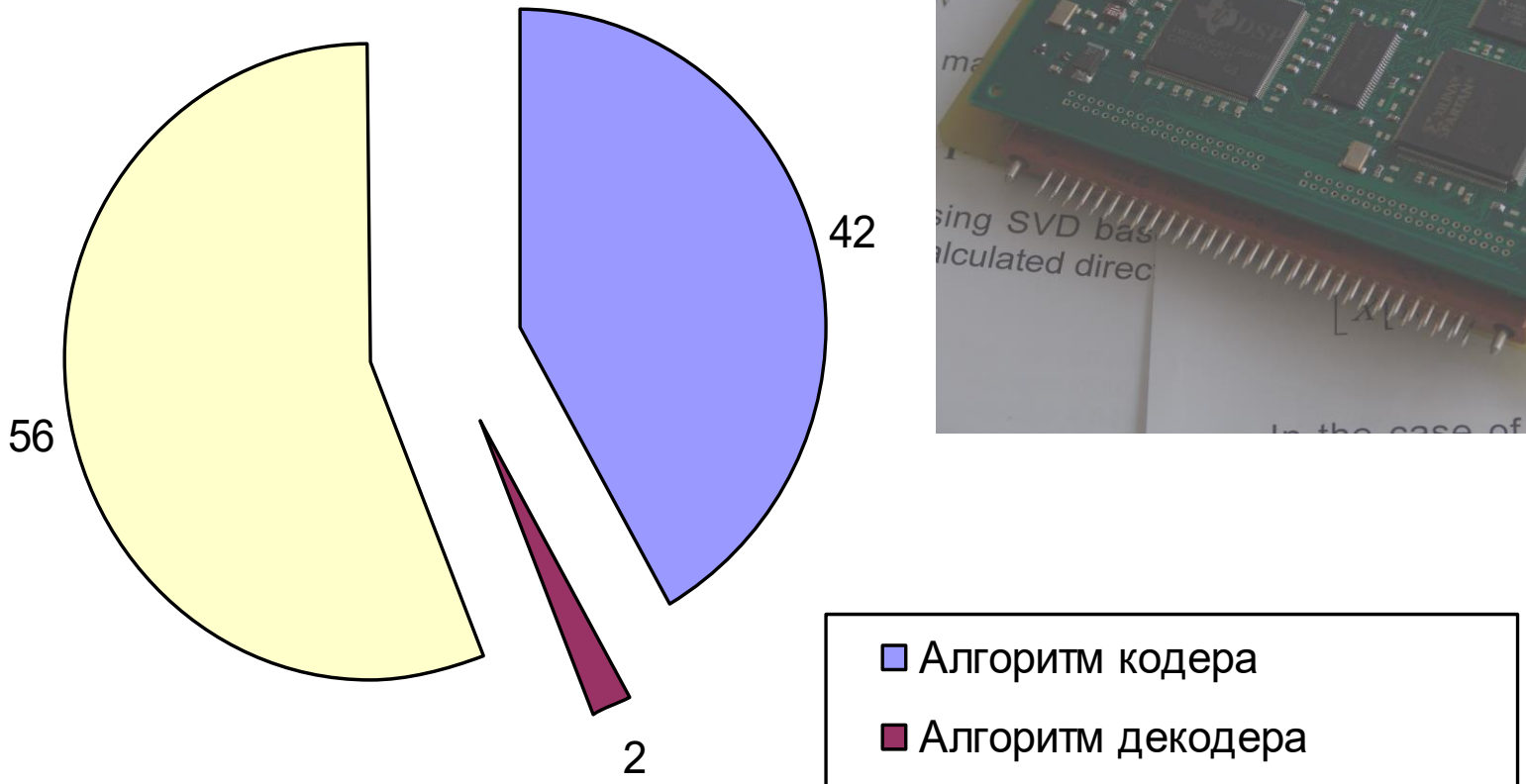


Лабораторный исследовательский комплекс на базе отладочной платы TMS320C6713 DSK



Гистограмма распределения вычислительного ресурса процессора TMS320C6713 (225 МГц) между алгоритмами кодера и декодера



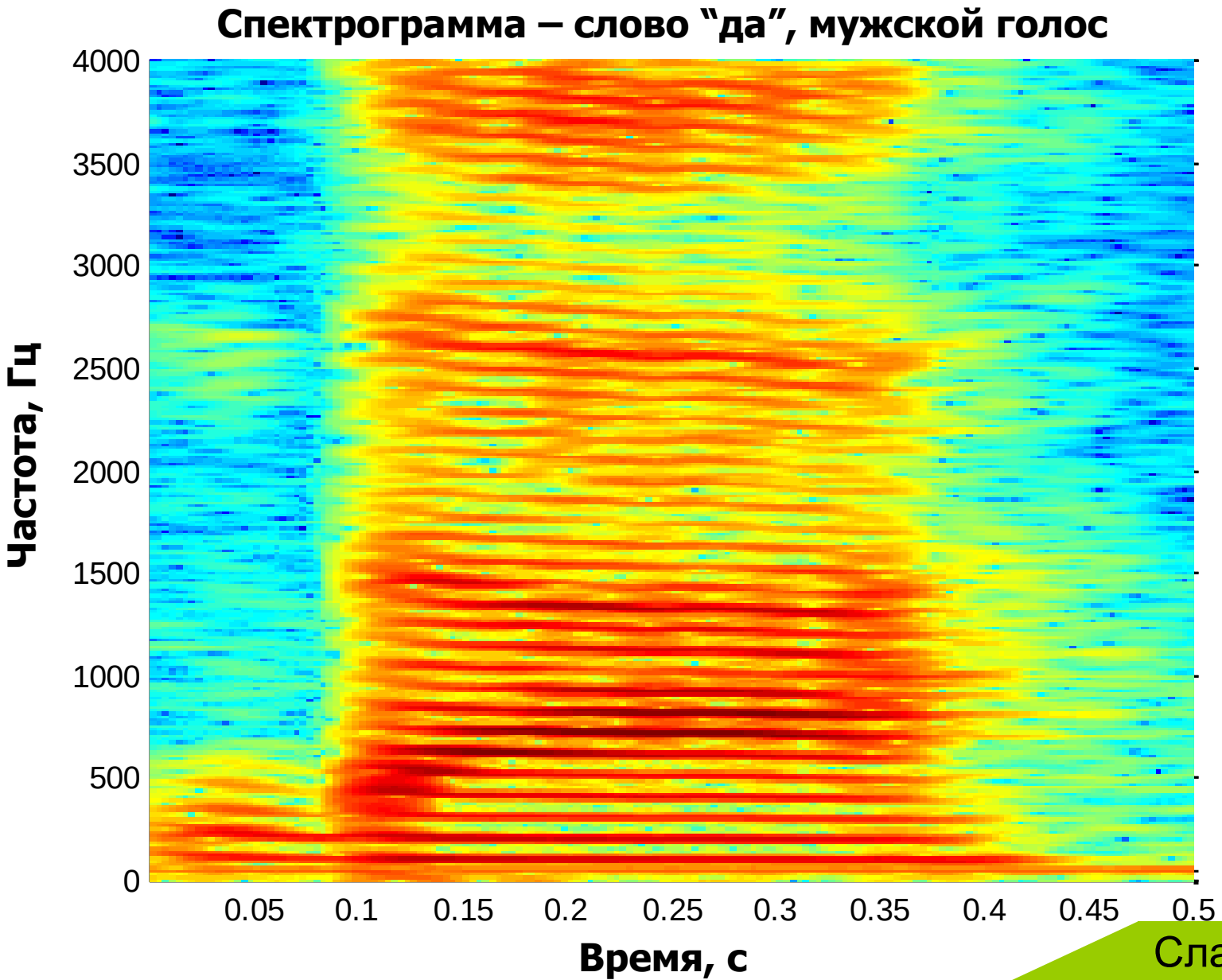


Субъективная оценка качества восстановленной речи

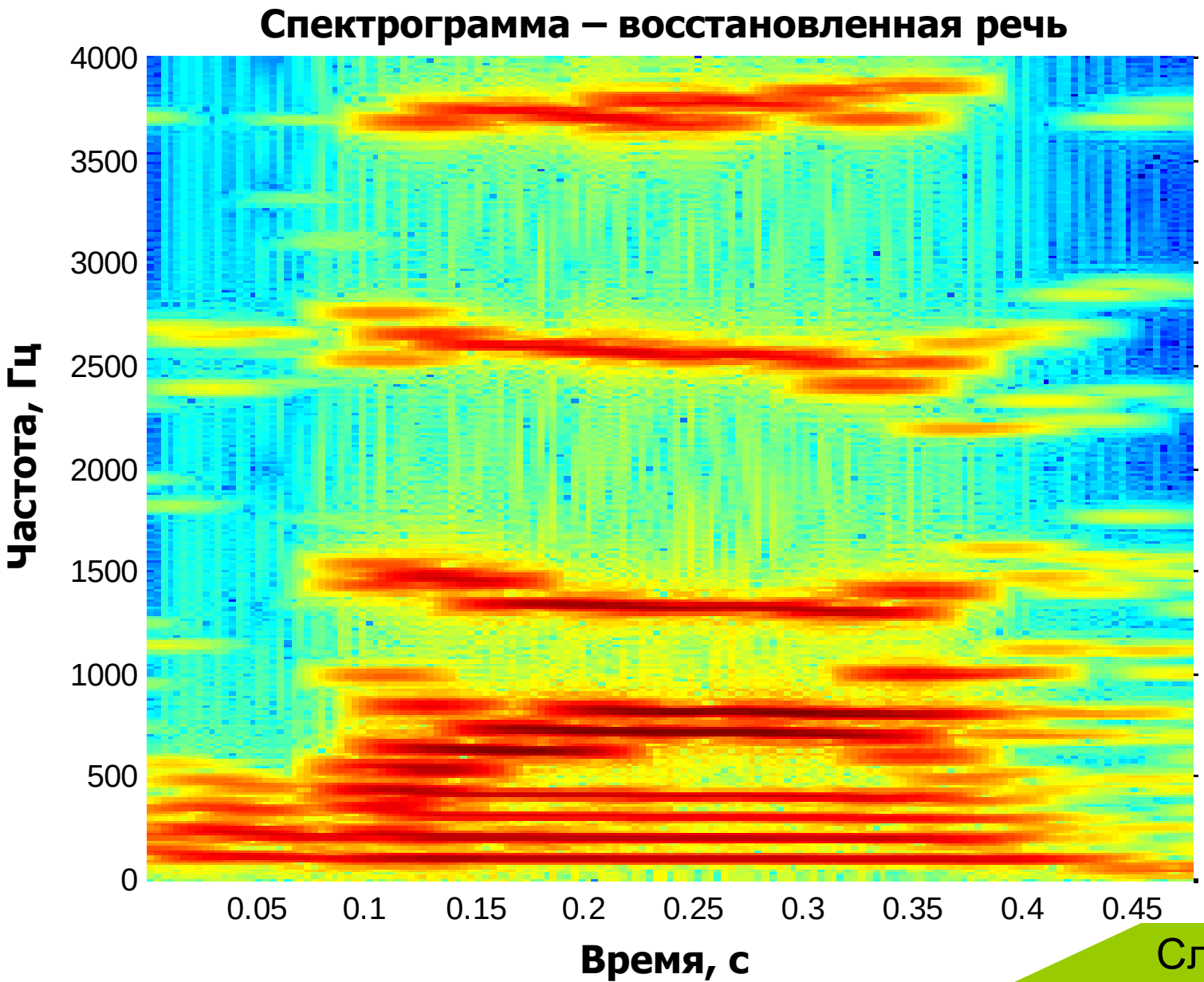
Скорость передачи, бит/с	Слоговая разборчи-вость речи, %	Качество, баллы	Характеристика восстановленной в декодере речи	Узнава-емость голоса диктора, %
8000	96	3.8	Некоторое нарушение естественности и узнаваемости, иногда присутствуют подзванивание и дребезжание	95

- Результаты субъективной оценки качества синтезированной речи подтверждают, что речь отличается довольно высокой степенью разборчивости и хорошей узнаваемостью диктора даже при ограниченном числе синусоидальных компонентов
- В соответствии с методикой из **СТБ ГОСТ Р 50840-2000** в ходе НИР «Разработать процедуры сжатия речевой информации, обеспечивающие коммерческое качество восстановленной речи», которая проводилась в рамках Государственной научно-технической программой «Развитие методов и средств системы комплексной защиты информации»

Экспериментальные результаты (1)



Экпериментальные результаты (2)



Основные результаты (1)

- На основании анализа существующих подходов синусоидального кодирования речевых сигналов применительно к системам компрессии было показано, что для достижения высокой степени компрессии **требуется введение антропоморфического критерия** для анализа входного сигнала. В ходе анализа практически доказана принципиальная возможность реализации системы компрессии речевого сигнала на основе синусоидальной модели.

- Разработан **новый метод антропоморфической обработки речевого сигнала** на основе трёх модифицированных моделей слуха человека: модели внешнего и среднего уха, комбинированной кохлеарной модели и модели слухового нерва, позволяющий производить отбор доминирующих частотных составляющих сигнала и обладающий на порядки меньшей вычислительной сложностью по сравнению с подобными алгоритмами.

- Предложен **метод реализации комбинированной кохлеарной модели в частотной области** путём совмещения пассивной и активной моделей, позволяющий более корректно обрабатывать речевые сигналы с различным уровнем. Для повышения точности кохлеарного анализа предложено использовать ДПФ с неравномерным частотным разрешением.

Основные результаты (2)

- Разработан **метод квантования синусоидальных параметров модели**, учитывающий особенности их общего статистического распределения и положения внутри набора параметров для каждого фрейма.

- Создано **алгоритмическое обеспечение** синусоидального вокодера с антропоморфической обработкой, особенностью которого является отсутствие процедур определения частоты основного тона и классификации фреймов, а при восстановлении речевого сигнала в декодере – использование небольшого числа синусоидальных компонент

- Осуществлена аппаратная реализация устройства компрессии-декомпрессии речевого сигнала. При использовании векторного квантования параметров и числе синусоидальных компонент не более 8 скорость потока данных составляет около 5 кбит/с. Экспериментальные результаты показывают, что речь отличается довольно высокой степенью разборчивости и хорошей узнаваемостью голоса диктора.