

# Речевые технологии. Введение.

- **Речь** — фундаментальный способ человеческой коммуникации.
- **Телефония и связи:** от классических звонков до VoIP (Zoom, Skype). Обеспечение разборчивости речи и подавление шума — базовая, но критически важная задача.
- **Медиаконтент:** музыка, фильмы, аудиокниги, подкасты.
- **Active noise control (ANC):** наушники в метро и самолёте, уменьшение шумов.
- **Реальное время:** трансляция и перевод, приложения, ломающие языковые барьеры.
- **Голосовые ассистенты и умные устройства:** Алиса, Маруся, Siri, Google Assistant.
- **Автоматизация общения:** голосовой набор, колл-центры, голосовые боты, анализ звонков.
- **Доступная среда:** синтез речи для незрячих, субтитры для слабослышащих.
- **Голосовая биометрия:** аутентификация и безопасность.
- **Медицина:** анализ голоса для диагностики, транскрибация диагнозов, удалённые консультации.
- **Неречевые звуки:** мониторинг, безопасность, контроль оборудования.
- **Автомобильные системы и умный дом.**

# Что такое звук

- **Звук — механическая продольная волна** колебаний частиц материальной среды (твёрдой, жидкой или газообразной).
- Вакуум **не проводит** звук: требуется среда для переноса импульса и энергии.
- Модель 1D: смещение частицы  $x(t)$ ; перепад давления  $p(t) = p_{\text{tot}}(t) - p_0$  распространяется со скоростью  $c$ .
- Геометрия волн: *плоская* (идеализация), *сферическая* (точечный источник), *цилиндрическая* (линейный источник).
- Продольность: колебания частиц направлены **вдоль** направления распространения волны.

# Основные параметры волны

- Амплитуда  $A$  — максимальное отклонение (смещения/давления) от среднего.
- Период  $T$  и частота  $f$ :  $T = \frac{1}{f}$ ,  $f$  в Гц.
- Длина волны  $\lambda$  — расстояние между соседними фазово-эквивалентными точками.
- Гармоническая волна:

$$s(x, t) = A \sin(2\pi f t - kx + \phi_0), \quad k = \frac{2\pi}{\lambda}$$

- Кинематическая связь через скорость звука:

$$c = \lambda f.$$

- Пример:  $f = 1 \text{ kHz}$  в воздухе ( $c \approx 343 \text{ m/s}$ )  $\Rightarrow \lambda \approx 0.343 \text{ м.}$

# Скорость звука и свойства среды

- В газах (адиабатически, идеальный газ):

$$c = \sqrt{\gamma RT} \quad (\text{прибл. } c \approx 331 + 0,6T_{\circ C} \text{ m/s})$$

- В жидкостях (К — модуль объёмной упругости):

$$c = \sqrt{\frac{K}{\rho}}$$

- В твёрдых телах (продольные волны, Е — модуль Юнга):

$$c = \sqrt{\frac{E}{\rho}}$$

- Сравнение (около 20 °C):

Воздух ~ 343 m/s

Вода ~ 1480 m/s

Сталь ~ 5900 m/s

# Звуковое давление: определение и единицы

- **Звуковое давление**  $p(t) = p_{\text{tot}}(t) - p_0$  — мгновенное отклонение от атмосферного  $p_0$ .
- Единицы: **Паскаль** (Pa). На практике используют *среднеквадратичное*  $p_{\text{rms}}$ .
- Диапазон характерных значений:

Порог слышимости (1 кГц)  $p_{\text{ref}} = 20 \mu\text{Pa}$

Разговор на 1 м  $\approx 0.02 \text{ Pa} (\approx 60 \text{ dB SPL})$

Автотрасса на 10 м  $\approx 0.2 \text{ Pa} (\approx 80 \text{ dB})$

Порог боли  $\approx 20 \text{ Pa} (\approx 120 \text{ dB})$

# Интенсивность и закон обратных квадратов

- **Интенсивность**  $I$  — поток акустической мощности через единицу площади ( $\text{W/m}^2$ ).
- Для плоской волны:

$$I = \frac{p_{\text{rms}}^2}{\rho c}, \quad Z = \rho c \text{ — акустический импеданс.}$$

- Точечный источник в свободном поле:

$$I(r) = \frac{P}{4\pi r^2}.$$

- Введем лог. меру интенсивности звука  $L_I = 10 \log_{10}\left(\frac{I}{I_0}\right)$  дБ  
где  $I_0 = 10^{-12} \text{ Вт/м}^2$  — минимально слышимая для человека.

# Логарифмическая мера интенсивности звука

- Для порога слышимости  $I = I_0$ , следовательно  $L = 0$
- Для сферической волны:

$$L_I(r_2) - L_I(r_1) = -20 \log_{10} \frac{r_2}{r_1} \text{ дБ},$$

т.е. при *удвоении расстояния*  $-6$  дБ, при *расстоянии  $10x$*   $-20$  дБ

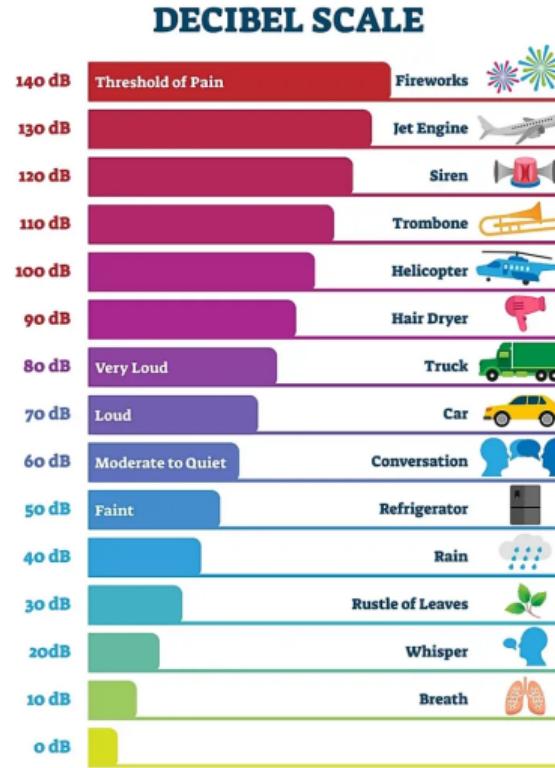
- Сложение некогерентных источников звука

$$L_{\text{tot}} = 10 \log_{10} \left( \sum_{i=1}^n 10^{L_i/10} \right)$$

т.е. для двух одинаковых  $+3$  дБ

- Для когерентных  $L_{\text{tot}} = 20 \log_{10} \left( \left| \sum_{i=1}^n 10^{L_i/20} e^{j\varphi_i} \right| \right)$  - в 2 раза больше некогерентных

# Сравнение лог. меры интенсивности звуков



# Отражение, поглощение и реверберация звука

## При встрече волны с преградой

- Баланс энергии на границе:  $R + A + T = 1$  (отражение, поглощение, прохождение).
- **Отражение:** зеркальное (угол падения = углу отражения) или диффузное (шероховатые поверхности рассеивают).
- **Поглощение:** коэффициент  $\alpha \in [0, 1]$ ; эквивалентная площадь поглощения  $A_e = \sum_i \alpha_i S_i$ .
- **Реверберация:** многократные отражения формируют «хвост» импульсной характеристики  $h(t)$ .

## Реверберационное время $RT_{60}$

- Определение: время убывания уровня звука на 60 dB после выключения источника.
- Оценка (Сабин):  $RT_{60} \approx \frac{0.161 V}{A_e}$ , где  $V$  — объём помещения ( $m^3$ ),  $A_e$  — эквив. площадь поглощения ( $m^2$ ).
- Речь: обычно целятся в 0.3 – 0.6 s для небольших комнат; > 1.5 s ухудшает разборчивость.

## Почему важно для ML по речи

- Сигнал в комнате:  $y(t) = s(t) * h(t) + n(t) \Rightarrow$  размывание временных и спектральных признаков.
- $s(t)$  - сигнал источника,  $h(t)$  - Room Impulse Response (RIR),  $n(t)$  - добавочный шум

## Как уменьшать $RT_{60}$

- Повышать  $A_e$ : материалы с высокой  $\alpha$ , ковры/шторы, акустические панели; избегать длинных параллельных «зеркальных» стен.

# Слух: диапазоны и уровни

- **Диапазон частот:** примерно 20 Гц–20 кГц; максимальная чувствительность около 2–5 кГц.
- **Уровень звукового давления (дБ SPL):**  $L_p = 20 \log_{10}(p/p_0)$ , где  $p_0 = 20 \mu\text{Па}$ .
- **Громкость** — субъективное ощущение: растёт не линейно с  $L_p$ ; прибавка  $\approx +10$  дБ воспринимается как примерно в 2 раза громче.
- **Динамический диапазон слуха:** от порога слышимости до болевого порога  $\sim 100$ –120 дБ.
- **Минимально заметная разница уровня:** порядка 1 дБ (около 1 кГц).

# Высота и частотное разрешение

- **Высота тона** связана с частотой: выше частота  $\Rightarrow$  выше ощущаемая высота.
- **Полосовой анализ уха:** слух разделяет спектр на *критические полосы* — диапазоны частот, в пределах которых сигналы сильно взаимодействуют.
- Ширина критических полос *растёт* с частотой (низкие частоты — узкие полосы, высокие — более широкие).
- Следствие: близкие по частоте компоненты на высоких частотах хуже различимы, чем на низких.

# Маскирование и время

- **Спектральное маскирование:** сильный звук может скрывать более слабый внутри той же критической полосы и по соседству.
- **Временное маскирование:** сильный звук может скрывать более слабый, звучащий *сразу после него* (вперёд по времени) и, слабее, *незадолго до него*.
- **Временная интеграция громкости:** энергия суммируется в окне порядка 100–200 мс.
- **Временная разборчивость:** изменения короче нескольких миллисекунд слышатся хуже.

# 1876 — Телефон Александра Белла

- Первая практическая передача речи по проводу.
- Старт эры электроакустики и телекоммуникаций.
- Базис для микрофонов и громкоговорителей.



Рис.: Телефон Белла

# 1877 — Фонограф Томаса Эдисона

- Первая запись и воспроизведение звука (цилиндр).
- Впервые можно хранить речь и музыку.



Рис.: Эдисон и фонограф

# 1887 — Граммофон Эмиля Берлинера

- Плоские диски вместо цилиндров ⇒ массовая запись.
- Стандартизация носителя и тиражирования.

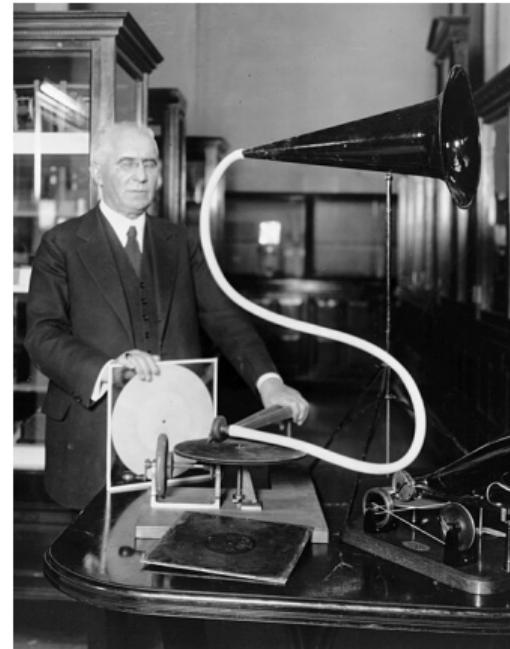


Рис.: Эмиль Берлинер с первым граммофоном

# 1898 — Время реверберации (Уоллес Сабин)

- Вводит понятие  $RT_{60}$  и формулу оценки реверберации.
- Рождение архитектурной акустики как науки.

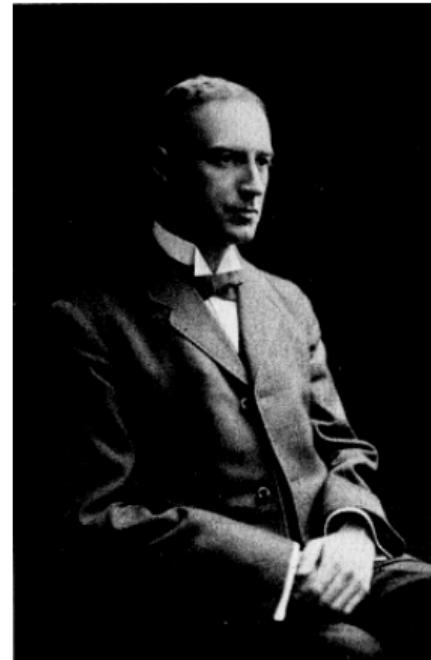


Рис.: Уоллес Сабин

# 1906 — Аудион (триод) Ли де Фореста

- Электронное усиление ⇒ качественная запись и радио.
- Существенный рост громкости и дальности связи.

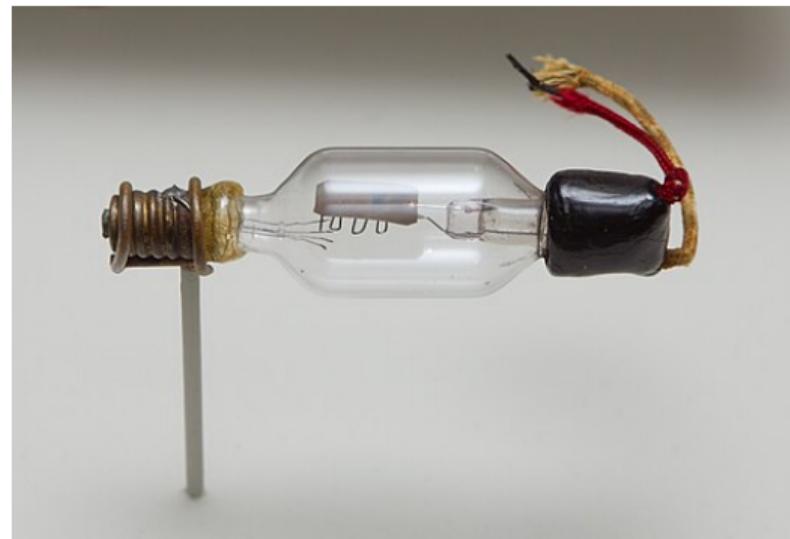


Рис.: Лампа «Audion» (триод), ранний усилитель.

# 1916 — Конденсаторный микрофон (Эдвард Уэнте)

- Высокая чувствительность и ровная амплитудно-частотная характеристика.
- Стандарт для качественной записи речи и музыки.
- Тонкая мембрана и неподвижная пластина образуют конденсатор, колебания воздуха изменяют его ёмкость,



Рис.: Western Electric 394

# 1928 — Магнитная лента

- Монтаж, перезапись, многодорожечность.
- Качественный и удобный носитель до цифры.
-  Play



Рис.: Magnetophon AEG 1934 и катушки с лентой.

# 1928/1948 — Теорема дискретизации Котельникова (Найквист–Шенон)

- Любую функцию  $F(t)$ , состоящую из частот от 0 до  $f_m$ , можно непрерывно передавать с любой точностью при помощи чисел, следующих друг за другом через интервалы времени  $\delta t = \frac{1}{2f_m}$  секунд
- 1897 Émile Borel
- 1915 E. T. Whittaker
- 1928 Harry Nyquist
- 1932 B. A. Котельников
- 1948 Claude Shannon

# 1937 — Импульсно-кодовая модуляция (PCM, Алекс Ривз)

- Цифровое представление звука: отсчёты и квантование.
- База для компакт-диска и цифрового аудио.

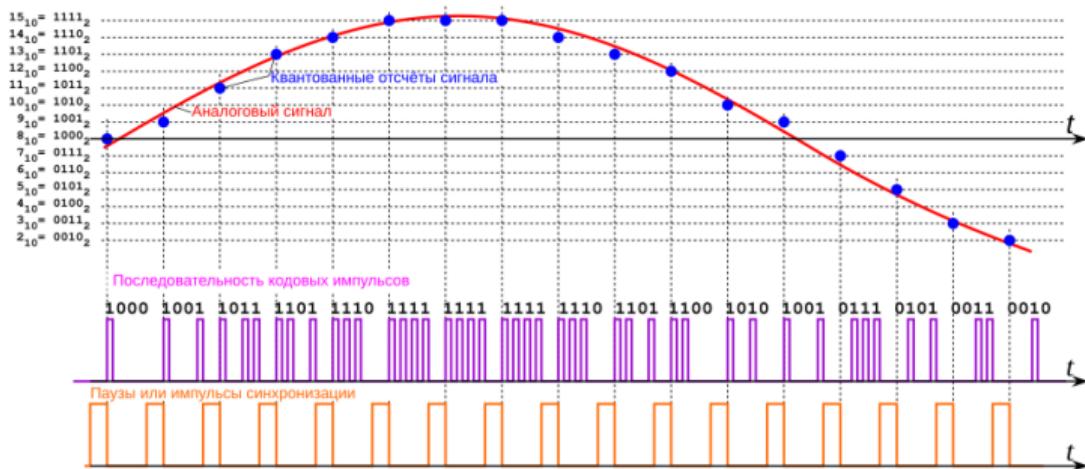


Рис.: Схема PCM. Пример 4-битной (16-уровневой) ИКМ. Показано квантование аналогового сигнала и пачки импульсов, кодирующих отсчёты. Передача в канале производится старшими битами вперёд.

# 1952 — «Audrey» (распознавание речи, Bell Labs)

- Распознавание десяти цифр по изолированной речи.
- Демонстрация принципиальной осуществимости ASR.
- Automatic Recognition of Spoken Digits K.H. Davis 1952

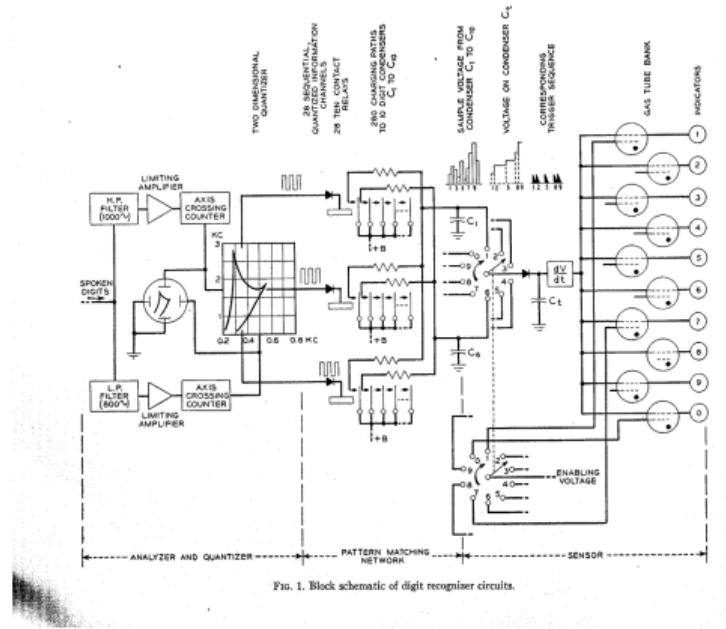


Рис.: Принципиальная схема Audrey.

# 1965 - Быстрое преобразование Фурье (FFT)

- **Определение.** FFT — семейство алгоритмов для быстрого вычисления дискретного преобразования Фурье (ДПФ) по данным длины N.

$$X_k = \sum_{m=0}^{N-1} x_m e^{-i2\pi km/N} \quad k = 0, \dots, N-1,$$

- **Сложность вычислений.** Наивно  $O(N^2)$ ; FFT  $O(N \log_2 N)$  — алгоритм Кули—Тьюки (1965); память  $O(N)$ .
- **Когда и кто.** 1965 — Дж. Кули и Дж. Тьюки (переоткрыли и популяризовали  $O(N \log N)$ ); ранние идеи — К. Ф. Гаусс (ок. 1805), Дэниелсон-Ланцош (1942).

# 1970-е — Линейное предсказание речи (LPC)

- Модель речевого тракта как фильтра автопредсказания.
- Компактные параметры для кодеков и анализа.
- Значение текущего отсчета речевого сигнала  $s(n)$  может быть предсказано как **линейная комбинация**  $p$  предыдущих отсчетов.

$$s(n) = \sum_{k=1}^p a_k s(n - k) + e(n)$$

$s(n)$  — текущий отсчет сигнала (то, что мы предсказываем).

$a_k$  — **коэффициенты предсказания** (LPC-коэффициенты). Это основная информация, которую мы извлекаем и храним.

$e(n)$  — **сигнал ошибки** (остаток). Разница между реальным и предсказанным сигналом.

- Кодируем по кадрам 20 мс, передаем LPC коэф., усиление, возбуждение. Типичный битрейт 2.4 кбит/с, возможно до 600 бит/с
- Используется в MPEG-4, OPUS, FLAC, LPC-10

# 1982 — Компакт-диск: 16 бит / 44.1 кГц (CD)

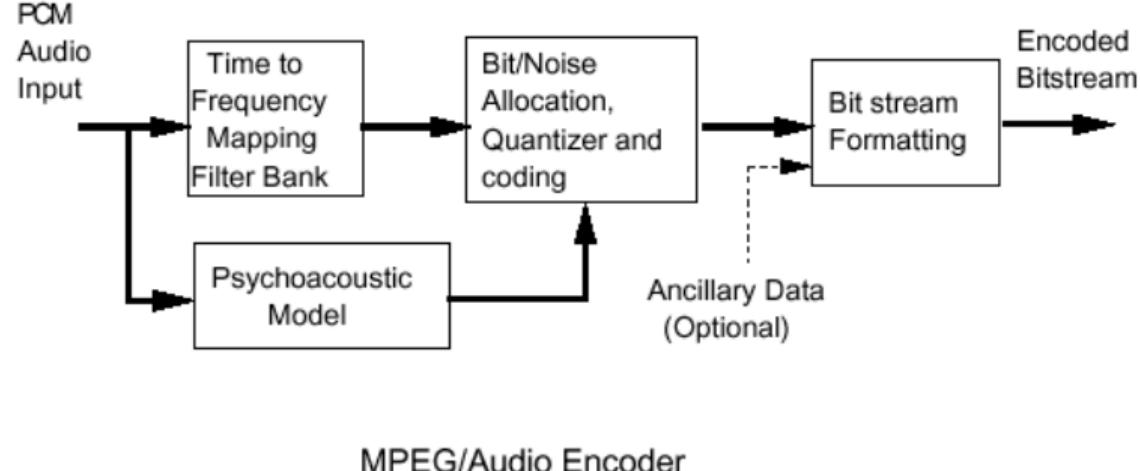
- Массовый цифровой звук на базе PCM.
- Стандарт качества и совместимости на десятилетия.



Рис.: Первый CD-диск

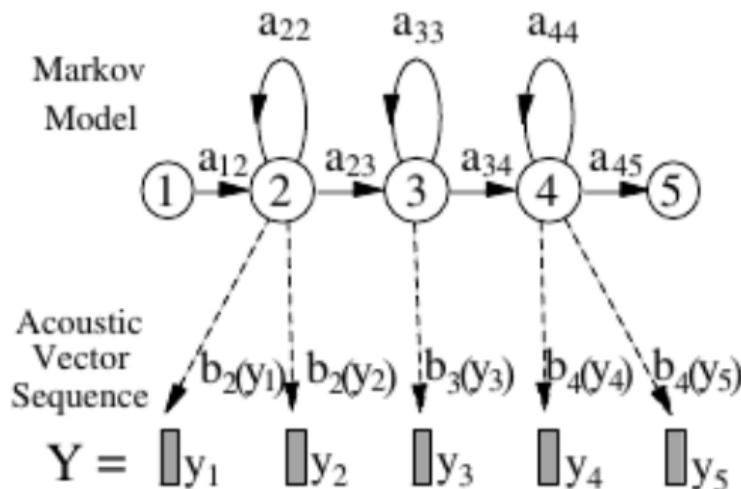
# Начало 1990-х — MP3 (перцептуальное кодирование)

- MPEG-1 Audio Layer III
- Использует маскирование слуха для экономии битрейта.
- Массовая цифровая дистрибуция аудио.



# 1970-1980-е — Скрытые марковские модели (НММ) в речи

- Вероятностная модель последовательностей для распознавания.
- Десятилетия промышленного стандарта до глубинных сетей.
- Каждое слово - своя модель, прогоняем через все модели, выбираем с максимальным правдоподобием.



# 2010-е — Глубокие нейросети для речи

- От гибридов «сеть + НММ» к сквозным моделям распознавания.
- LSTM, свертки, CTC
- Трансформеры, attention.

# 2020-е — настоящее время

- Conformer 2020
- Самообучение на аудио (предобучение без разметки) SSL. Модели учатся по неразмеченному звуку, затем дообучаются.
- Рост робастности к шуму, реверберации и акцентам.
- Модели ASR вышли на уровень человека

# Mel-спектrogramma: определение

- Mel-спектrogramma — матрица энергий спектра, спроектированная с линейной оси частот (Гц) на психоакустическую шкалу mel.
- Элемент  $S_{\text{mel}}[i, n]$  — энергия  $n$ -го фрейма в  $i$ -й mel-полосе.
- Предназначена для согласования частотного разрешения со слуховым восприятием: выше детализация на НЧ, грубее на ВЧ.
- Очень часто используется в ML моделях как вход звука

# Обозначения и параметры

- Сигнал:  $x[t]$ ,  $t = 0, \dots, T - 1$ , частота дискретизации  $f_s$ .
- Длина окна  $L$  (сэмплы), шаг  $H$  (сэмплы), обычно  $L \approx 20\text{--}40$  мс,  $H \approx 10$  мс.
- Длина БПФ  $N_{FFT} \geq L$  (часто степень 2).
- Число полос  $M$  (напр., 40, 64, 80).
- Диапазон:  $[f_{min}, f_{max}]$ , обычно  $f_{min} \in [0, 50]$  Гц,  $f_{max} = f_s/2$ .
- (Опц.) pre-emphasis:  $y[t] = x[t] - \alpha x[t - 1]$ ,  $\alpha \in [0.95, 0.98]$ . Часто  $\alpha = 0.97$

# Окноное разбиение (STFT)

- STFT = Short-Time Fourier Transform — кратковременное преобразование Фурье (КВПФ).
- Фрейм  $n$ :  $s_n[m] = y[nH + m] \cdot w[m]$ ,  $m = 0, \dots, L - 1$ .
- Окна:
  - Ханна:  $w_{\text{Hann}}[m] = \frac{1}{2} \left(1 - \cos \frac{2\pi m}{L-1}\right)$ .
  - Хэмминга:  $w_{\text{Hamming}}[m] = 0.54 - 0.46 \cos\left(\frac{2\pi m}{L-1}\right)$ .
  - Блэкмана:  $w_{\text{Blackman}}[m] = 0.42 - 0.5 \cos\left(\frac{2\pi m}{L-1}\right) + 0.08 \cos\left(\frac{4\pi m}{L-1}\right)$ .
- Нулевое дополнение до  $N_{\text{FFT}}$  по необходимости.

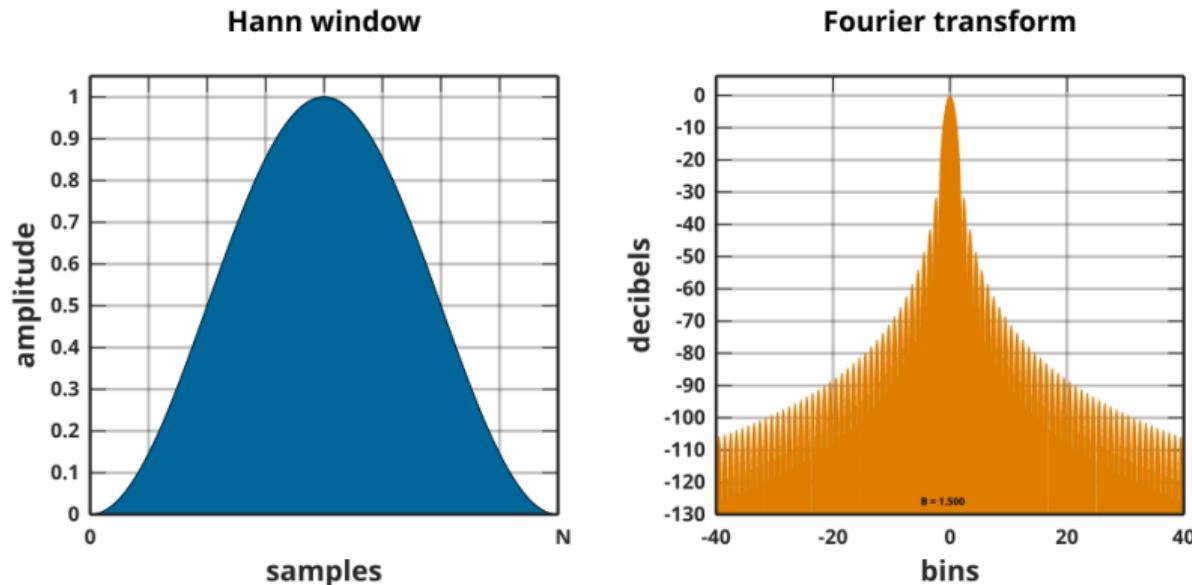
# Теорема о свёртке (Фурье-преобразование свёртки)

- Пусть  $f, g \in L^1(\mathbb{R})$ . Свёртка:  $(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau$ .
- Нормировка Фурье:  $F(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$ ,  
 $f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega$ .
- Тогда:  $\mathcal{F}\{f * g\}(\omega) = F(\omega) G(\omega)$ .
- Эквивалентно:  $\mathcal{F}^{-1}\{F(\omega)G(\omega)\}(t) = \frac{1}{2\pi} (f * g)(t)$ .
- Следствие (двойственность):  $\mathcal{F}\{f \cdot g\}(\omega) = \frac{1}{2\pi} (F * G)(\omega)$ .
- Замечание о нормировке: при другой нормировке коэффициенты с  $2\pi$  меняются соответствующим образом.

# Окно Ханна (Hann)

- Формула:

$$w[n] = 0.5 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1.$$

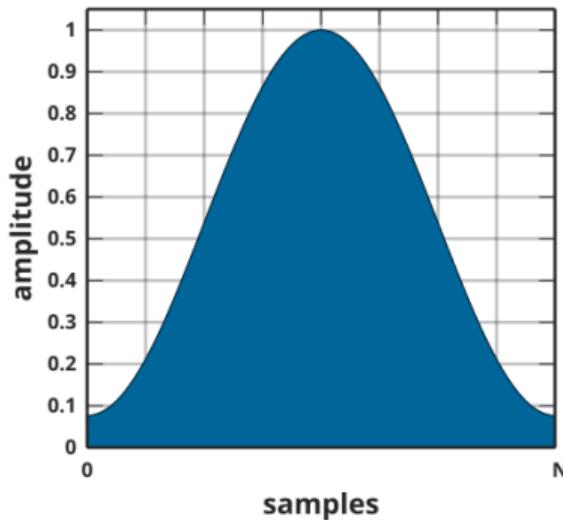


# Окно Хэмминга (Hamming)

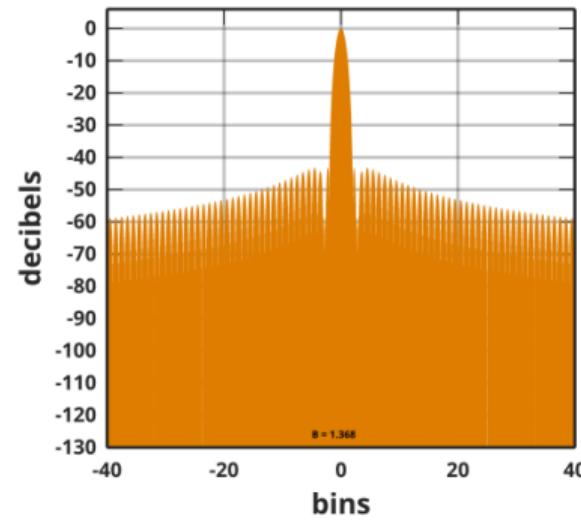
- Формула:

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1.$$

Hamming window ( $a_0 = 0.53836$ )



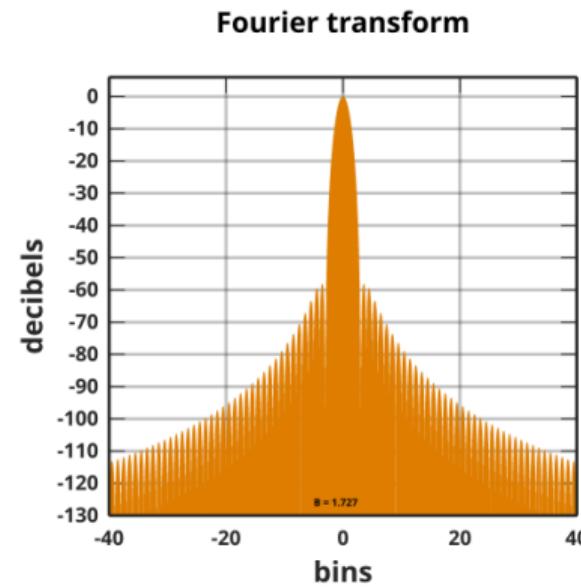
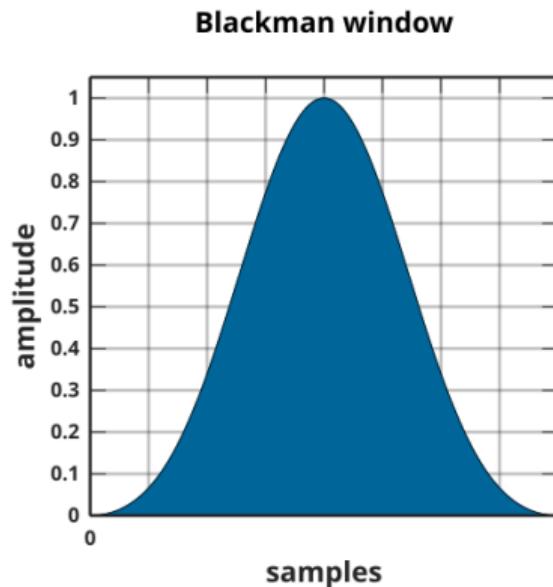
Fourier transform



# Окно Блэкмана (Blackman)

- Формула:

$$w[n] = 0.42 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cos\left(\frac{4\pi n}{N-1}\right), \quad 0 \leq n \leq N-1.$$



# БПФ и спектральная мощность

- БПФ:

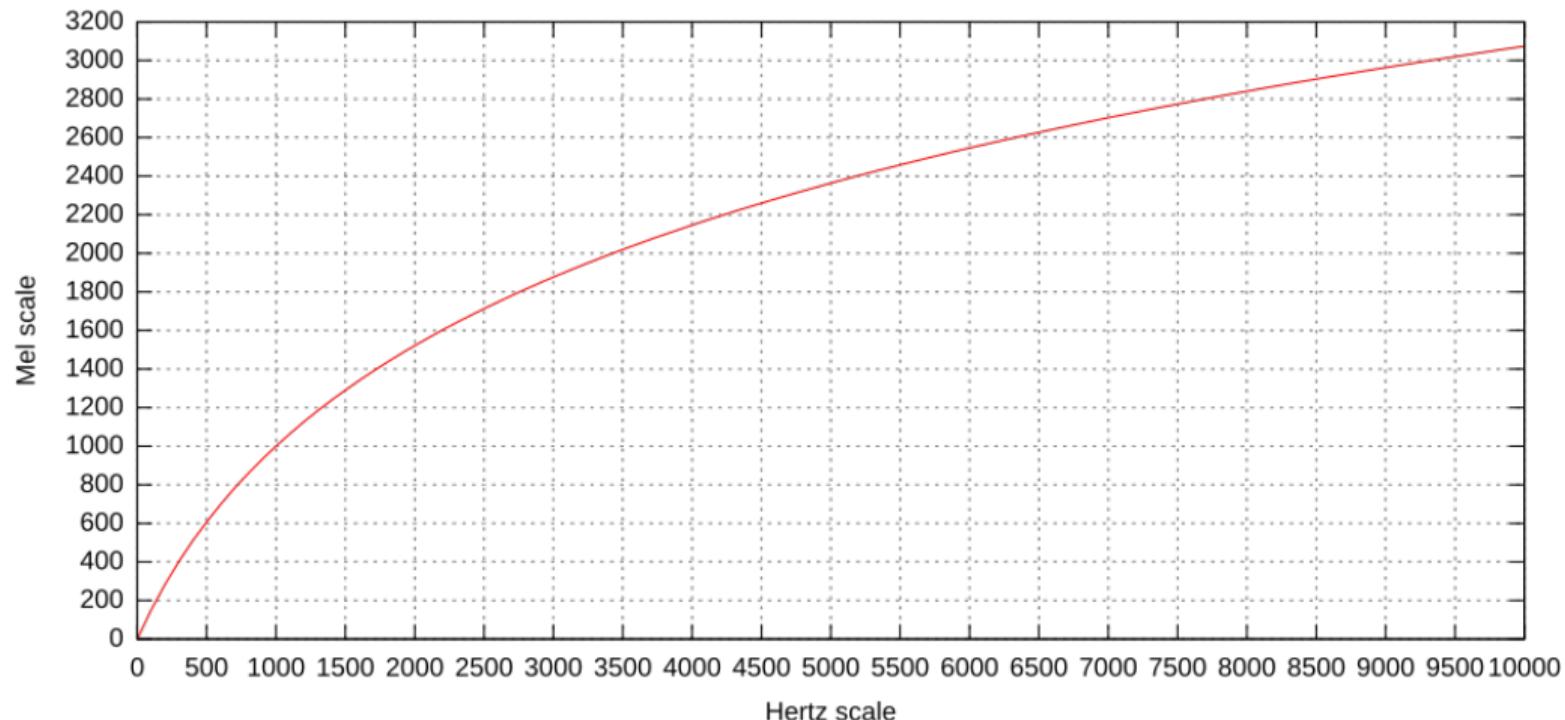
$$X_n[k] = \sum_{m=0}^{N_{\text{FFT}}-1} s_n[m] e^{-j2\pi km/N_{\text{FFT}}}$$

- Односторонний спектр:  $k = 0, \dots, K$ ,  $K = \lfloor N_{\text{FFT}}/2 \rfloor$ .
- Биновая частота:  $f_k = \frac{k f_s}{N_{\text{FFT}}}$ .
- Мощность:  $P_n[k] = \frac{1}{N_{\text{FFT}}} |X_n[k]|^2$  (или амплитуда  $A_n[k] = |X_n[k]|$ ).

# Шкала mel (формулы)

- Прямой переход:  $m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$ .
- Обратный переход:  $f(m) = 700(10^{m/2595} - 1)$ .
- Эквивалент через  $\ln$ :  $m(f) = 1127 \ln\left(1 + \frac{f}{700}\right)$ ,  $f(m) = 700(e^{m/1127} - 1)$ .

# Шкала mel



# Узлы фильтробанка (на мелах)

- Границы в мел:  $m_{\min} = m(f_{\min})$ ,  $m_{\max} = m(f_{\max})$ .
- Узлы:  $m_i = m_{\min} + i \frac{m_{\max} - m_{\min}}{M + 1}$ ,  $i = 0, \dots, M + 1$ .
- Обратно в Гц:  $f_i = f(m_i)$ , получаем  $\{f_0, \dots, f_{M+1}\}$ .

# Треугольные mel-фильтры

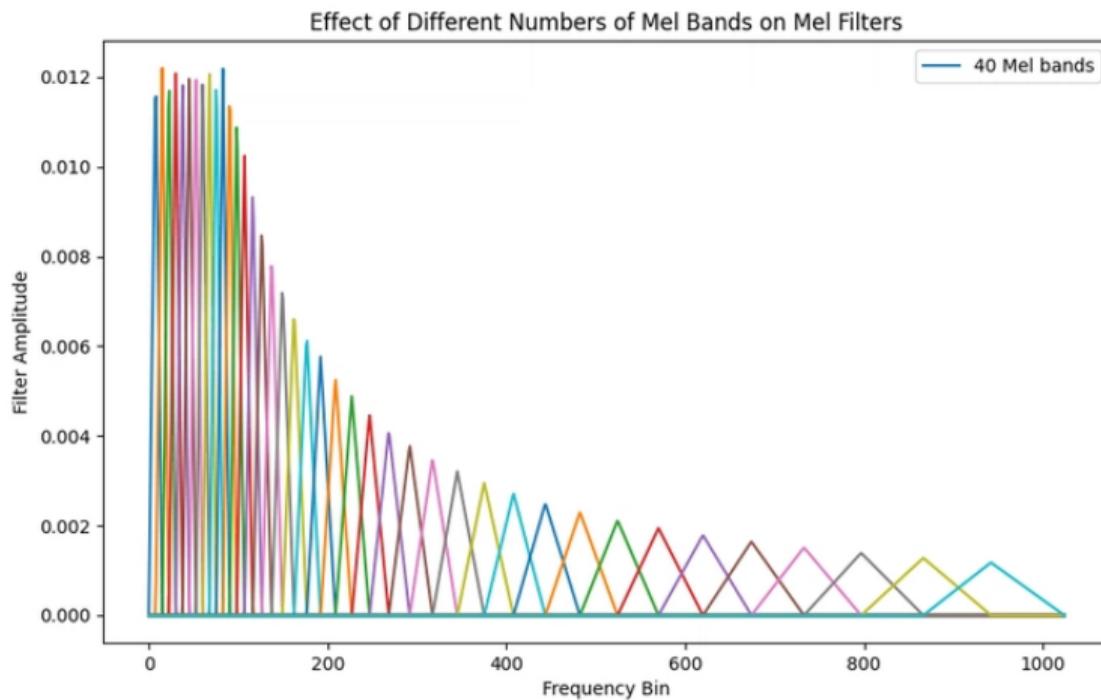
- Для  $i = 1, \dots, M$  с опорами  $(f_{i-1}, f_i, f_{i+1})$  и бином  $f_k$ :

- 

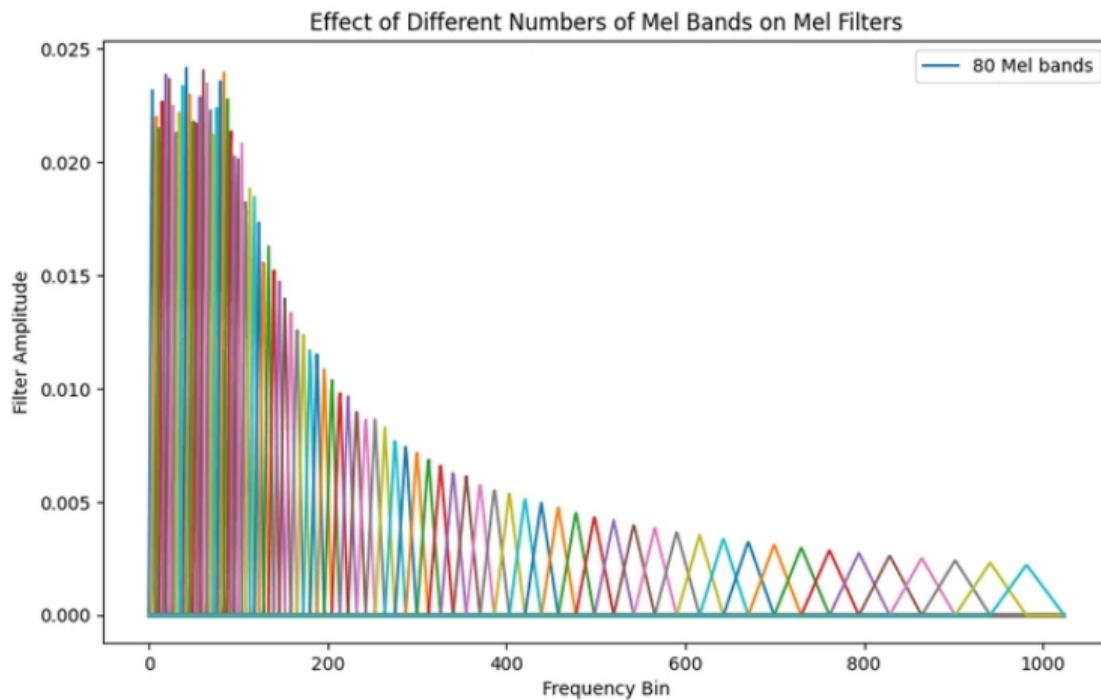
$$h_i[k] = \begin{cases} 0, & f_k < f_{i-1} \text{ ИЛИ } f_k > f_{i+1}, \\ \frac{f_k - f_{i-1}}{f_i - f_{i-1}}, & f_{i-1} \leq f_k \leq f_i, \\ \frac{f_{i+1} - f_k}{f_{i+1} - f_i}, & f_i \leq f_k \leq f_{i+1}. \end{cases}$$

- HTK-нормализация (**Hidden Markov Model Toolkit**): вершина = 1 (как выше).
- Slaney-нормализация (площадь  $\approx 1$ ):  $\tilde{h}_i[k] = \frac{2}{f_{i+1} - f_{i-1}} h_i[k]$ .

# Треугольные mel-фильтры



# Треугольные mel-фильтры



# Проекция мощности на mel-полосы

- Энергия полосы  $i$  во фрейме  $n$ :  $S_{\text{mel}}[i, n] = \sum_{k=0}^K h_i[k] P_n[k]$ .
- Матричная форма: если  $H \in \mathbb{R}^{M \times (K+1)}$ ,  $P_n \in \mathbb{R}^{K+1}$ ,
- то  $S_{\text{mel}}[:, n] = H P_n$ .

# Логарифм и dB-представление

- Лог-масштаб:  $S_{\log}[i, n] = \ln(S_{mel}[i, n] + \varepsilon)$ ,  $\varepsilon \sim 10^{-10}$ .
- Децибелы (мощность):  $S_{dB}[i, n] = 10 \log_{10} \frac{S_{mel}[i, n]}{p_{ref}}$ .
- Для амплитуды использовать  $20 \log_{10}(\cdot)$  вместо  $10 \log_{10}(\cdot)$ .
- Опорный уровень  $p_{ref}$ : 1.0 или максимум по  $S_{mel}$  (нужно зафиксировать для воспроизводимости).

# Минимальный алгоритм (шаги)

- Задать  $f_s$ , L, H,  $N_{FFT}$ , M,  $f_{min}$ ,  $f_{max}$ ,  $\varepsilon$ .
- Построить окно  $w[m]$  (Ханна, Хэмминга, Блэкмана).
- Для каждого фрейма: применить окно, нулевое дополнение, БПФ, взять  $k = 0..K$ .
- Вычислить мощность  $P_n[k] = |X_n[k]|^2/N_{FFT}$  (или амплитуду  $A_n[k] = |X_n[k]|$ ).
- Выбрать нормализацию НТК или Slaney
- Один раз построить матрицу фильтробанка H по  $\{f_i\}$  и  $f_k$ .
- Получить  $S_{mel} = H \cdot P$  для всех фреймов.
- Применить  $\ln$  или  $10 \log_{10}$  к  $S_{mel}$  (или  $20 \log_{10}$  к амплитуде).

- Типичные значения:  $L = 25$  мс,  $H = 10$  мс;  $N_{FFT} \in \{512, 1024, 2048\}$ .
- Полосы  $M \in \{40, 64, 80\}$ ; для моделей TTS/ASR часто 80.
- $f_{min} = 20\text{--}30$  Гц для речи;  $f_{max} = \min(8000, f_s/2)$  при ограничении речевого диапазона.
- Pre-emphasis улучшает ВЧ-доли речи; может быть отключено для музыки/нейросетей.
- Масштаб окна и одно-сторонняя энергия: придерживаться выбранной конвенции во всей цепочке.

# Совместимость и «острые углы»

- Уточнять: формулы  $\text{mel} \leftrightarrow \text{Hz}$  (основание логарифма).
- Выбор нормализации фильтров: HTK vs Slaney; фиксировать и документировать.
- Power vs magnitude; соответствующая формула dB (10 vs 20).
- Эпсилон  $\varepsilon$  для  $\ln$ : фиксировать значение для детерминизма.
- Диапазон  $[f_{\min}, f_{\max}]$ : влияет на первые/последние полосы и энергетику.
- Сложность:  $O(N_{\text{frames}} N_{\text{FFT}} \log N_{\text{FFT}} + M K)$ .

# Примеры спектрограмм

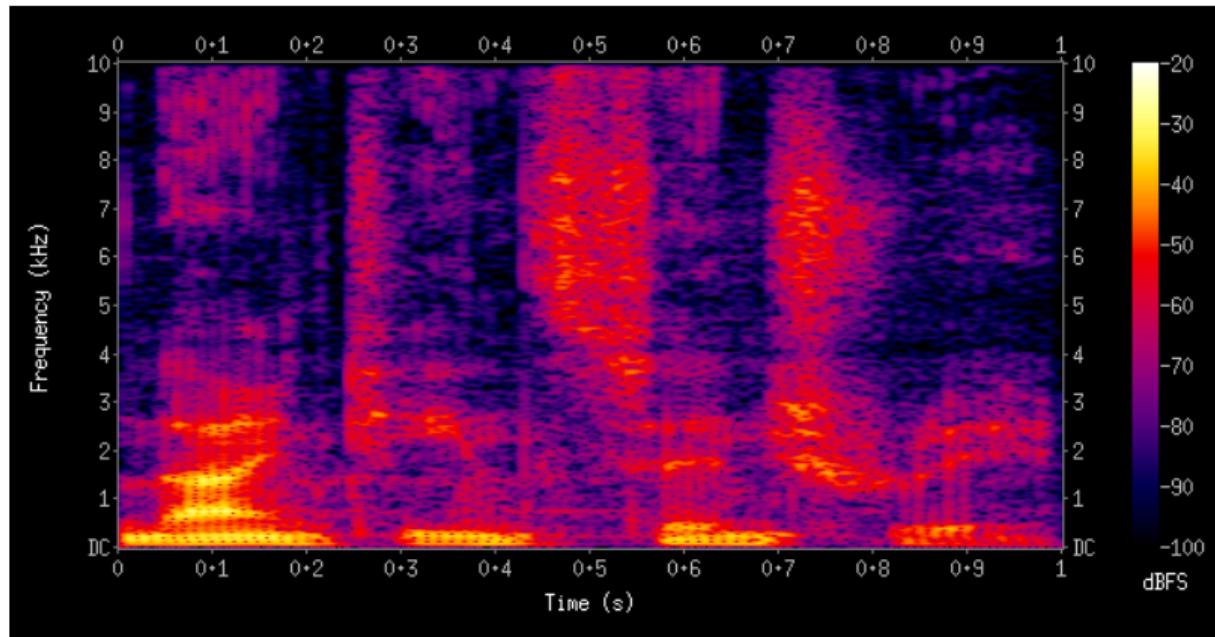


Рис.: Мужской голос говорит "nineteenth century"

# Примеры спектрограмм

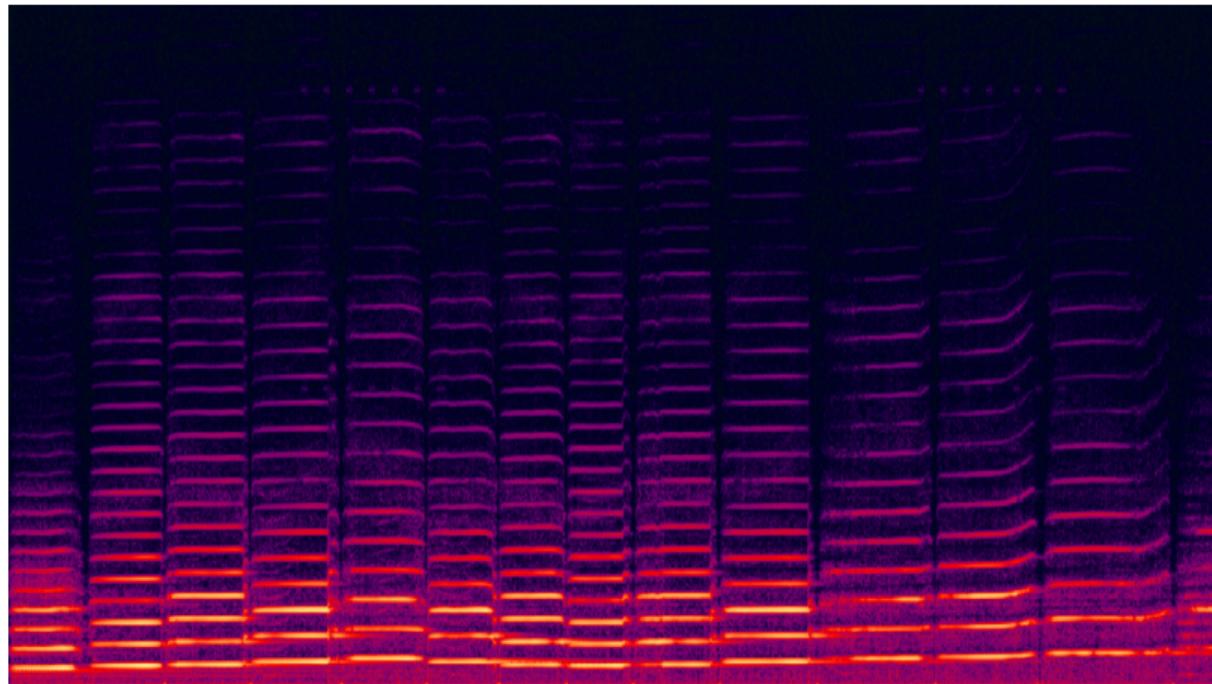


Рис.: Игра на скрипке

# Примеры спектрограмм

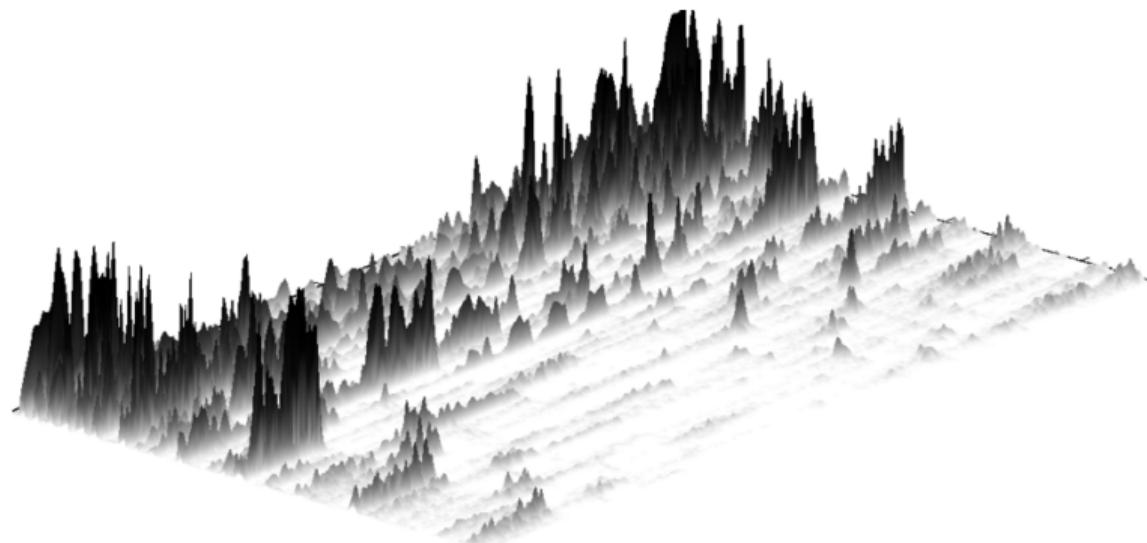


Рис.: 3D представление спектrogramмы классического произведения

# I Saved a PNG Image To A Bird

- Один блоггер записал PNG в скворце
- <https://www.youtube.com/watch?v=hCQCP-5g5bo>



# I Saved a PNG Image To A Bird

