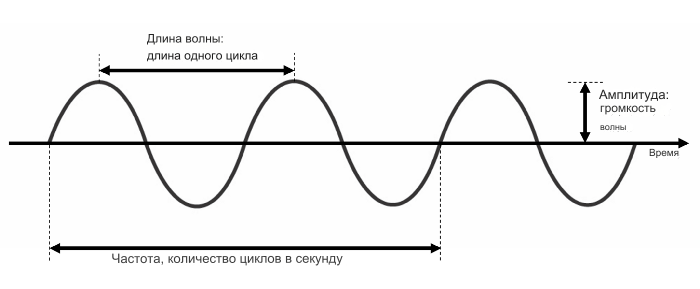
**Мел-спектрограмма 2**

Перемещение звуковых волн во времени требует больших вычислительных ресурсов, и мы можем получить больше информации из двумерных данных изображений, чем из одномерных волн. Кроме того, недавнее быстрое развитие компьютерного зрения, особенно с помощью свёрточных нейронной сетей, может значительно улучшить результаты, если рассматривать аудио как изображения.

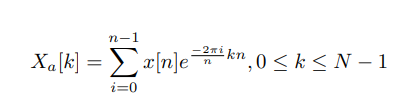
Звуки могут быть представлены в форме волн, и волны имеют два важных свойства: частоту и амплитуду, как показано на рисунке. Частота определяет высоту звука, амплитуда — его громкость.



Спектрограмма – это представление сигнала в виде спектров, полученных на каждом временном отрезке анализируемого сигнала.

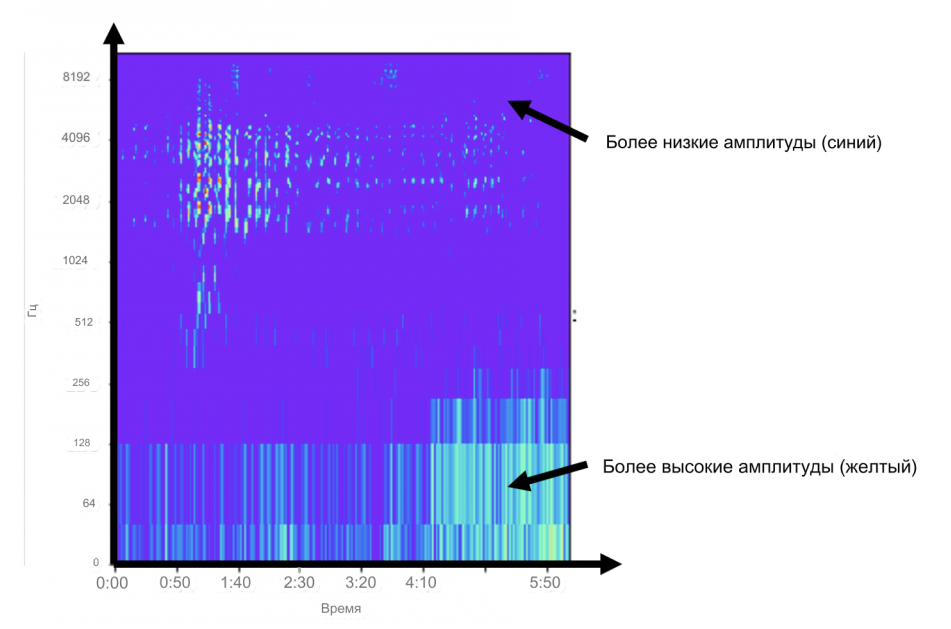
Спектр – разложение аудиосигнала по гармоникам. Гармоника – это частоты кратные самой низкой частоте исходного сигнала. У спектрограммы на оси абсцисс отложено время, а на оси ординат частота, на самой спектрограмме откладывается спектр в разные моменты времени. Третье измерение, показывающее изменение амплитуды, отмечается интенсивностью цвета. С помощью спектрограмм задача обработки звука может быть представлена как задача обработки изображений.

Для построения спектрограммы используется преобразование Фурье:

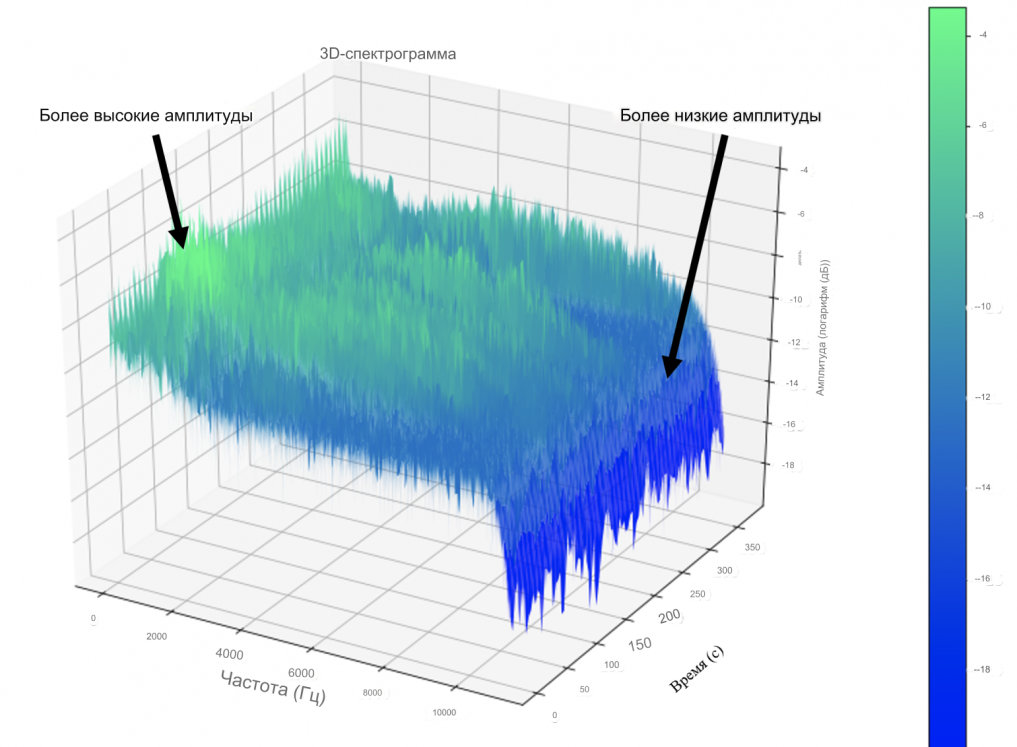


Преобразование Фурье разлагает сигнал на составляющие его частоты и отображает амплитуду каждой частоты, присутствующей в сигнале. На практике используется кратковременное преобразование Фурье, которое разбивает звуковой сигнал на кадры, так как в обычном преобразовании невозможно распознать в какой момент времени было конкретное распределение частот. Из временных кадров формируется спектрограмма.

На спектрограмме аудиоклипа горизонтальное направление представляет время, а вертикальное направление представляет частоты. Наконец, амплитуда звуков определённой частоты, существующих в определённый момент, представлена цветом точки, который обозначается соответствующими координатами x-y.



Чтобы понять, как частоты отражаются в спектрограммах, посмотрите на трёхмерную визуализацию, которая демонстрирует амплитуду с помощью дополнительного измерения. По оси X отложено время, а по оси Y — значения частот. Ось z — это амплитуда звуков частоты координаты y в момент координаты x. По мере увеличения значения z цвет меняется с синего на красный, получается цвет, который мы видели в предыдущем примере 2D-спектрограммы.



Спектрограммы полезны, потому что они показывают именно ту информацию, которая нам нужна: частоты, особенности, которые формируют форму слышимого нами звука.

Но если использовать спектрограммы, полученные таким образом, сигнал на них будет практически незаметен или сосредоточен в одной области. Это происходит из-за того, что человек иначе воспринимает

звук – человеческий слух не воспринимает различия во всех частотных диапазонах одинаково. По мере увеличения частот нам становится всё труднее различать их. Чтобы лучше имитировать поведение человеческого уха с помощью моделей глубокого обучения, мы измеряем частоты по шкале мел. В шкале мел любое равное расстояние между частотами звучит для человеческого уха одинаково. Единица мел (m) связана с герцами (f) таким уравнением:



У нас также есть свои особенности восприятия высоты тона. Низкие частоты для нас звучат «низко», а высокие - «высоко», но связь между частотой и восприятием нелинейна - синусоидальные волны 100 и 200 Гц будут звучать гораздо дальше друг от друга, чем волны 10 000 и 10 100 Гц.

Попробуем построить спектрограмму и мел-спектрограмму сами.

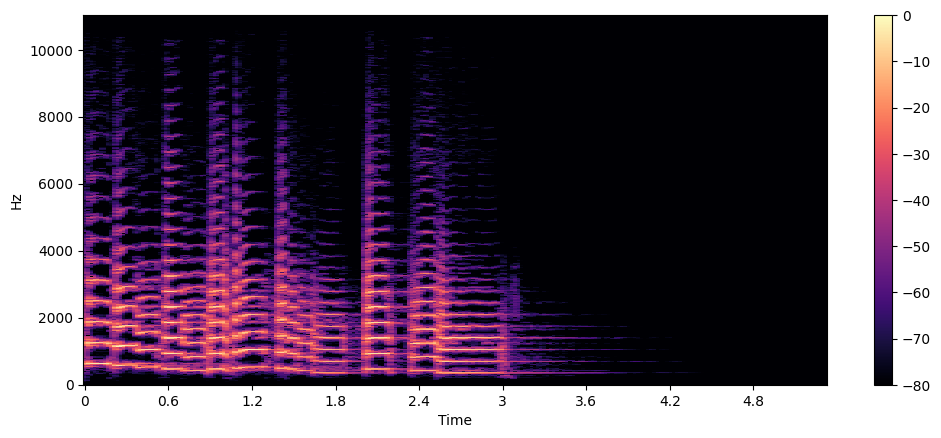
Импортируем аудиофайл с записью трубы из библиотеки librosa:

| import librosa # Функция load возвращает два значения: array (массив аудиоданных) и sampling\_rate (частота дискретизации аудиофайла). array, sampling\_rate = librosa.load(librosa.ex("trumpet")) |
| --- |

[audio.wav](https://drive.google.com/file/d/13oYuBaaNbllGHLHqdgmCFGJATqsSBn8-/view?usp=sharing)

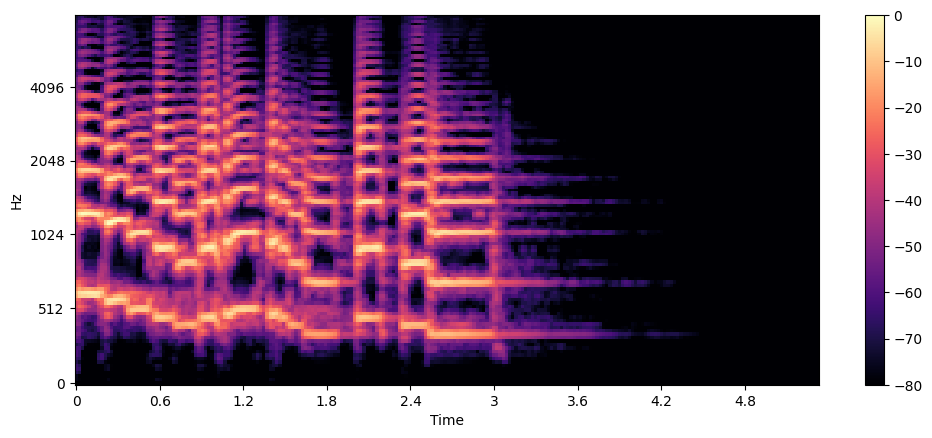
Сгенерируем спектрограмму:

| import numpy as np import matplotlib.pyplot as plt import librosa.display # Вычисляем коротко-временное преобразование Фурье (STFT) для массива аудиоданных D = librosa.stft(array) # Преобразуем амплитуду STFT в децибелы S\_db = librosa.amplitude\_to\_db(np.abs(D), ref=np.max)  plt.figure().set\_figwidth(12) # Отображаем спектрограмму с использованием librosa.display.specshow # Указываем x\_axis="time" для отображения времени по оси x и y\_axis="hz" для отображения частоты по оси y librosa.display.specshow(S\_db, x\_axis="time", y\_axis="hz") # Добавляем цветовую шкалу к спектрограмме plt.colorbar() |
| --- |



Сгенерируем мел-спектрограмму:

| # Вычисляем мел-спектрограмму для массива аудиоданных S = librosa.feature.melspectrogram(y=array, sr=sampling\_rate, n\_mels=128, fmax=8000) # Преобразуем мощность спектрограммы в децибелы S\_dB = librosa.power\_to\_db(S, ref=np.max)  # Создаем новую фигуру для построения мел-спектрограммы plt.figure().set\_figwidth(12) # Указываем x\_axis="time" для отображения времени по оси x и y\_axis="mel" для отображения мел-шкалы по оси y # Устанавливаем sr=sampling\_rate и fmax=8000 для корректного масштабирования осей librosa.display.specshow(S\_dB, x\_axis="time", y\_axis="mel", sr=sampling\_rate, fmax=8000) plt.colorbar()  Преобразуем мощность спектрограммы в децибелы S\_dB = librosa.power\_to\_db(S, ref=np.max)  # Создаем новую фигуру для построения мел-спектрограммы plt.figure().set\_figwidth(12) # Указываем x\_axis="time" для отображения времени по оси x и y\_axis="mel" для отображения мел-шкалы по оси y # Устанавливаем sr=sampling\_rate и fmax=8000 для корректного масштабирования осей librosa.display.specshow(S\_dB, x\_axis="time", y\_axis="mel", sr=sampling\_rate, fmax=8000) plt.colorbar() |
| --- |



Мы видим на изображении, что светлые фрагменты, обозначающие высокий уровень энергии звука стали крупнее, а их отличие по размеру друг от друга с увеличением частоты стало меньше.

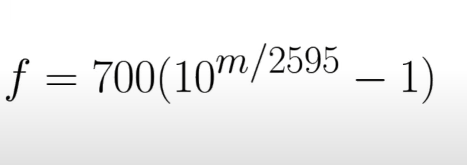
| restored\_audio = librosa.feature.inverse.mel\_to\_audio(S)  Audio(data=restored\_audio, rate=sampling\_rate) |
| --- |

[restored\_audio.wav](https://drive.google.com/file/d/1BDupay0MOJBf-OtKAvNsSX1AYJxAHF-j/view?usp=sharing)

Слышно, что при восстановлении звука из мел-спектрограммы, полученное аудио немного искаженное. Также видно, что выделенные фрагменты.

Здесь мы преобразовали мел-спектрограмму обратно в линейную шкалу.

Наша исходная мел-спектрограмма имеет логарифмический масштаб по частоте. Обратное преобразование требует отмены этого логарифмического преобразования, чтобы вернуться к линейной шкале частот. Это можно сделать с помощью обратной функции отображения мел-шкалы в герцы:



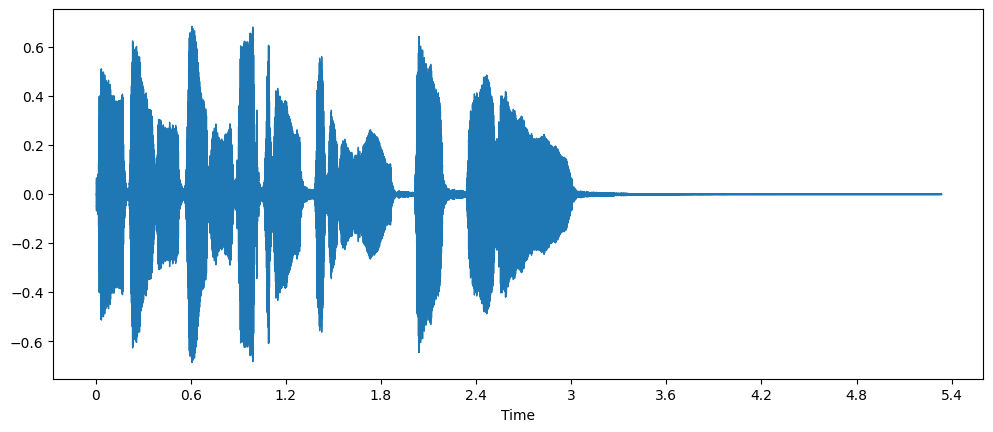
Затем спектрограмма амплитуд преобразуется обратно в временной сигнал. Это часто делается с использованием обратного коротко-временного преобразования Фурье (ISTFT):



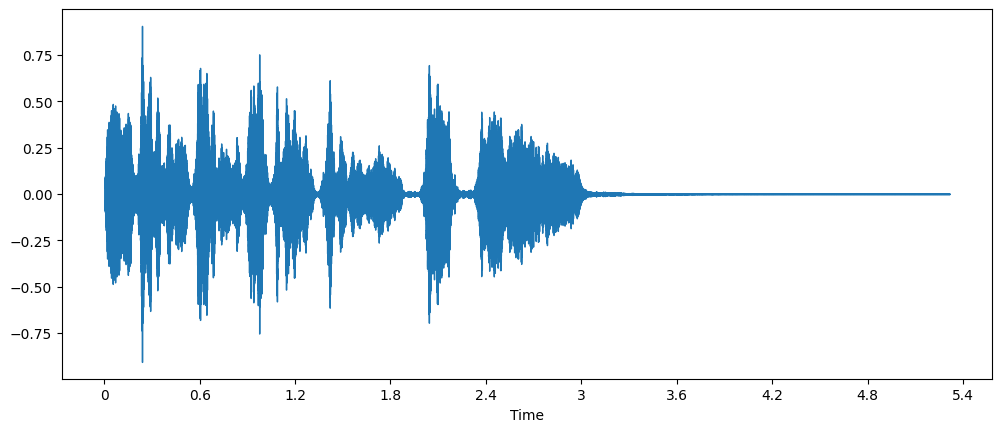
Оно принимает спектрограмму амплитуд и фазовую информацию, чтобы восстановить временной сигнал. Полученный временной сигнал может быть восстановлен и сохранен в аудиофайле, который звучит аналогично исходному аудиофайлу.

Сравним формы звуковой волны и частотный спектр аудиосигнала изначального аудио и восстановленного из мел-спектрограммы.

Audio

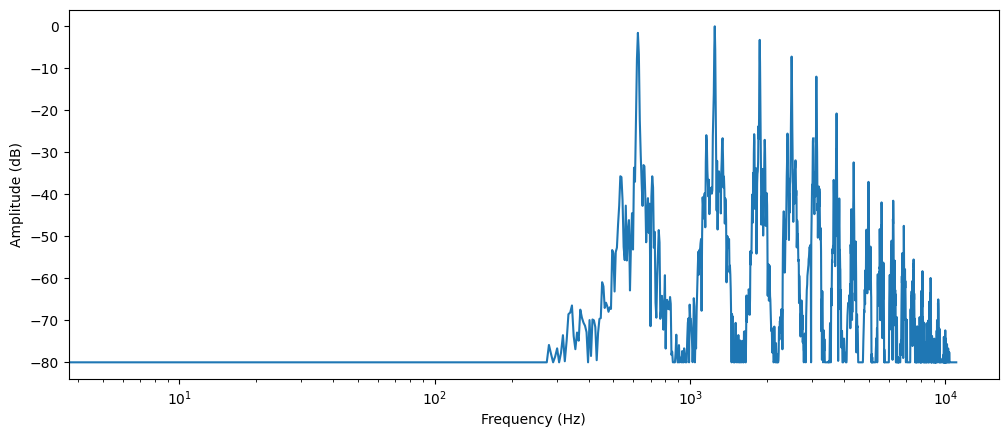


Restored\_Audio:

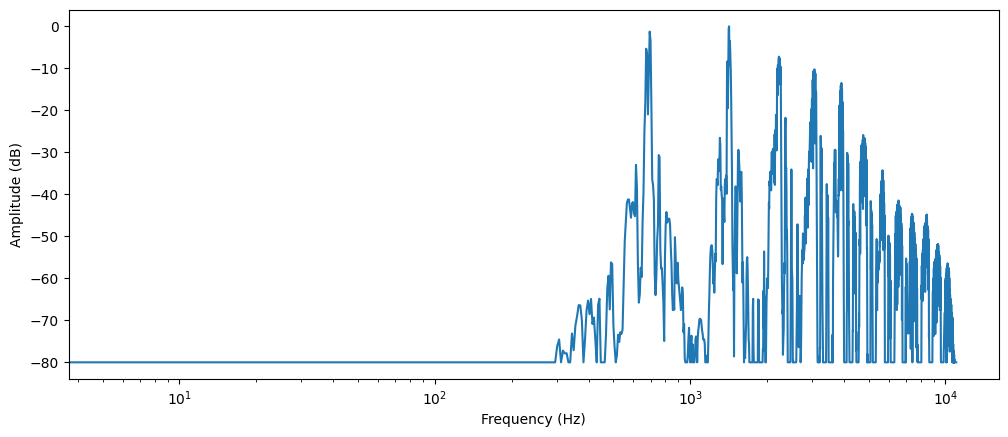


В восстановленном аудио более частый перепад амплитуд, но общее изображение почти такое же.

Audio:



Restored audio:

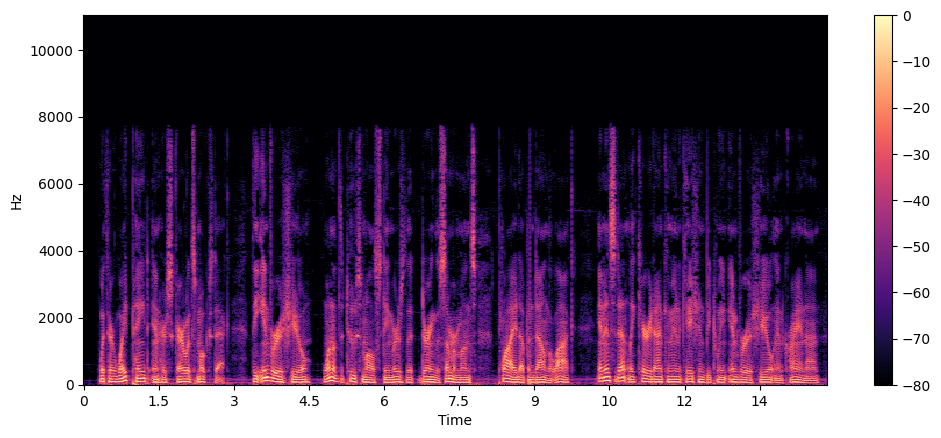


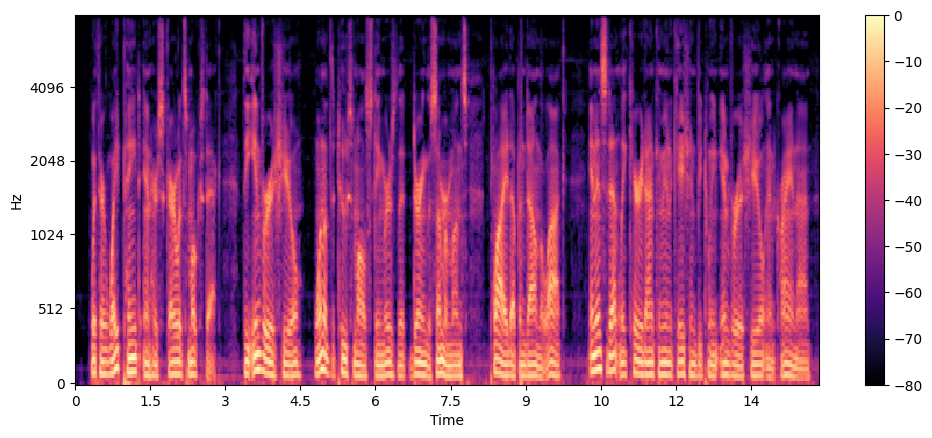
Аналогичная ситуация наблюдается и для частотного спектра.

Попробуем теперь повторить ту же процедуру, но с образцом человеческой речи. Для этого используем 14-секундный отрывок с мужским голосом из библиотеки librosa.

[human\_speech.wav](https://drive.google.com/file/d/1IusUig9e87GTZKIZ78hEuuQmBdMegctF/view?usp=sharing)

Спектрограмма:



Мел-спектрограмма:  


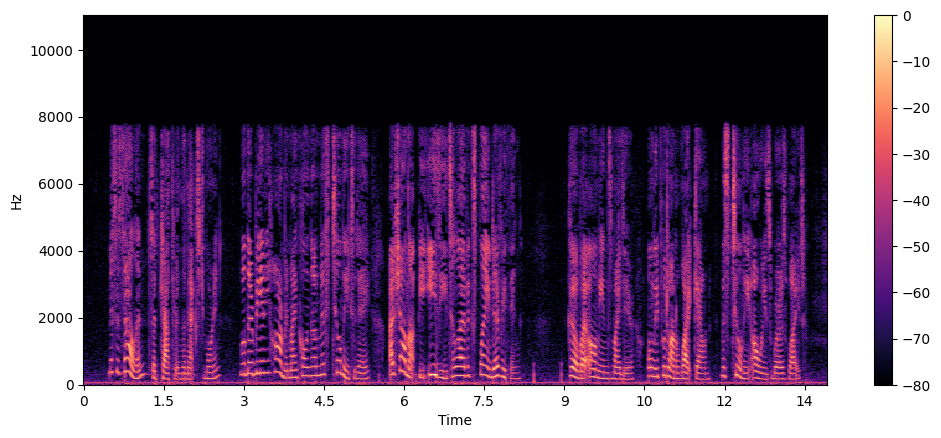
Восстановленное аудио из мел-спектрограммы:

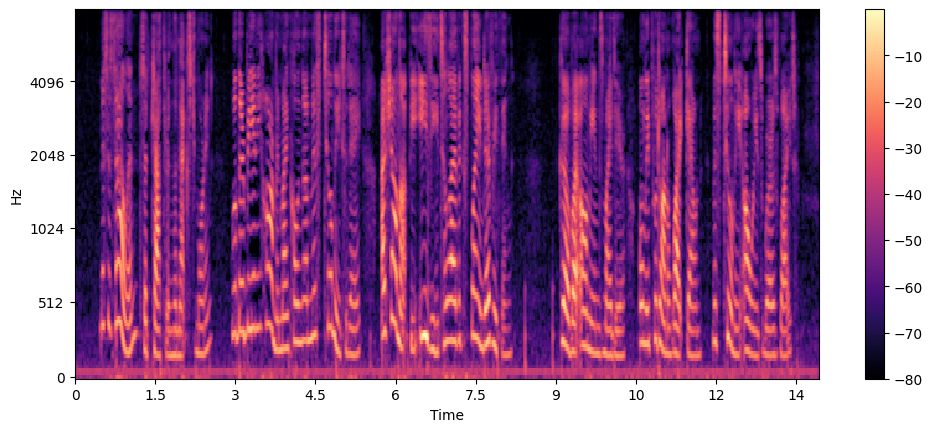
[re\_human\_speech.wav](https://drive.google.com/file/d/1pxJpC-OLa2x8x4zXgb6aGULxR1ONlKht/view?usp=sharing)

Возьмем теперь образец женской речи:

[wooman\_speech.wav](https://drive.google.com/file/d/1iX5PB8CpHIW1EkAw4bexWkTks7VV8d9a/view?usp=sharing)

Спектрограмма:



Мел-спектрограмма:  


Восстановленное аудио из мел-спектрограммы:

[re\_woman\_speech.wav](https://drive.google.com/file/d/10Bnxl1qmTBti2gXDc5gEUMkQPiUhjuNq/view?usp=sharing)

Вывод, который можем сделать: Паузы и тишина выделена темным цветом, речь выделяется светлыми участками. Мужчина говорит на более низких тонах и монотонно, поэтому светлее всего выделяется нижняя часть спектрограммы. Женский голос более высокий и в данном случае менее монотонный, поэтому светлые участки на мел-спектрограмме расположены выше и нет скопления в одном месте, как в мужской речи.

При переводе в мел-шкалу происходит уплотнение частот, особенно в низких частотах, что может привести к потере частотной точности, могут быть потеряны некоторые высокие частоты.

Мел-спектрограмма обычно ограничена динамическим диапазоном значений, что может привести к потере информации в тихих или очень громких участках аудиосигнала.

Из-за особенностей мел-шкалы могут быть упрощены детали в высоких частотах, что может привести к потере информации о высокочастотных компонентах сигнала.

Весь код:

[mel-spectrogram.ipynb](https://colab.research.google.com/drive/1oSvyCVgVXCCfL2P6FZyV27oNQqS8-z1u?usp=sharing)

Источники:

<https://cyberleninka.ru/article/n/metody-augmentatsii-audio-signala/viewer>

<https://habr.com/ru/companies/skillfactory/articles/536834/>

<https://huggingface.co/learn/audio-course/ru/chapter1/audio_data>

<https://www.soundsandwords.io/audio-loss-functions/>

<https://www.youtube.com/watch?v=9GHCiiDLHQ4>