***Фурье***

Для обучения и дальнейшего использования модели для распознавания речи необходимо определить способ представления и передачи звука для дальнейшей его обработки нейросетью. Для глубокого обучения нейросетей, работающих со звуком, как правило, используются файлы формата WAV (англ. waveform — «в форме волны»). Рассмотрим подробнее устройство файлов данного формата. В области данных WAV-файла хранятся последовательные значения амплитуды звука в определенный момент времени. Продолжительность промежутка времени, соответствующего одному отсчету в секундах, определяется величиной, называемой частота дискретизации и равен 1/SR, где SR – частота дискретизации (Hz). В [теореме Котельникова](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1%80%D0%B5%D0%BC%D0%B0_%D0%9A%D0%BE%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%B8%D0%BA%D0%BE%D0%B2%D0%B0) доказано, что для идеального воссоздания исходной звуковой волны таким образом достаточно использовать частоту дискретизации, вдвое превышающую самую высокую частоту записываемого звука. Типичным значениями частоты дискретизации является 44100 Hz, однако для распознавания речи достаточно и 16000 Hz, так как диапазон частот человеческой речи не столь велик.

Таким образом, звуковой моно-сигнал для обработки нейросетью с помощью WAV-файлов можно представить в виде одномерного массива. Такое представление мало информативно и не применимо для адекватного решения задачи распознавания речи. Также значительной проблемой данного представления является размер полученного одномерного массива. Количество точек в нем будет равно SR \* T, где SR – частота дискретизации, D - продолжительность аудиофайла в секундах. Так, например, при частоте дискретизации равной 44100 Hz массив, представляющий запись длиной 10 секунд, будет содержать 441 тысячу точек. Очевидно, что предоставлять данные такого объема на вход нейросети не эффективно.

Для приведения полученного массива точек к более информативному виду, который используется в большинстве систем распознавания речи на данный момент, его преобразуют в спектрограмму. Рассмотрим процесс данного преобразования. Основной задачей является получение распределения влияния гармонических колебаний определенной частоты звука на представленную в виде массива амплитуд звуковую волну. Эту задачу можно решить с помощью преобразования Фурье по формуле:

g~(f) = int (-inf - inf) g(t) e^(-2 pi i t f) dt,

где f – частота,

t – время.

Таким образом из функции g(t): R->R зависимости амплитуды от времени, получаем функцию g~(f): R->C, отражающую зависимость коэффициентов Фурье от частоты звука. Это предположение верно для непрерывного времени и частоты от минус бесконечности до бесконечности. В случае работы с одномерными массивами действительных чисел время дискретно и частоты принимают ограниченные значения. Как уже известно количество точек в массиве равно N = Sr \* D, пусть тогда количество отсчетов частоты также равно N. Это поможет сохранить обратимость преобразования между функциями g~ и g. Тогда будем использовать дискретное преобразование Фурье:

g~(k\*Sr/N-1) = sum (0 - N-1) g(n) e^(-2 pi I / N \* kn),

где n - дискретные временные точки n = 0\_N-1

k - индекс частоты, k = 0\_N-1. (при этом частота равна f = k\*Sr/N-1, и принимает значения от 0 до Sr).

Таким образом из функции g(t): [0, N]->R зависимости амплитуды от времени, получаем функцию g~(f): [0, Sr]->C, отражающую зависимость коэффициентов Фурье от частоты звука. Взяв модуль от коэффициентов Фурье для каждой частоты меньшей, чем Sr/2, так как именно эти частоты дискредизируются без потерь, получим нужное нам представление.

Однако подобное преобразование неэффективно на длинных аудиофайлах, и, в частности, для задач распознавания речи, ведь в данных задачах требуется учитывать подобное распределение не на протяжении всей записи, а в каждый конкретный момент времени. Для решения этой задачи используется оконное преобразование Фурье. Разобьем изначальный массив на отрезки, каждый из которых содержит M точек. Тогда можно применить дискретное преобразование Фурье к каждому из этих отрезков и получить распределение для времени, которое им соответствует. Основной проблемой данного подхода является растекание спектра. Растекание спектра (от англ. «spectral leakage») – это явление, из-за которого измеренная спектральная энергия перетекает из одной частоты на другие частоты. Это происходит, когда дискретизированный сигнал содержит нецелое число периодов за время измерения. Искусственное прерывание сигнала, которое появляется при разбиении его на отрезки, проявляется в виде высокочастотных составляющих, которых не было в исходном сигнале. Для борьбы с растеканием спектра применяют более гладкую оконную функцию. И преобразование Фурье производится над произведением оконной функции и исходного сигнала. Обычно используется окно Ханна:

w(n) = 1/2 (1- cos(2 pi n / N-1))

где *N* — ширина окна.

Очевидно, что, если перемножать исходный сигнал и оконную функцию на последовательных отрезках, информация о сигнале в начале и конце каждого отрезка теряется, потому что сигнал там умножается на числа, близкие к нулю. Для избежание этой проблемы в оконном преобразовании Фурье отрезки берут не последовательно, а совмещая конец предыдущего отрезка с началом следующего.

В таком случае, формула для оконного преобразования Фурье выглядит следующим образом:

S (m, k\*Sr/N) = sum (0 - N-1) g(n + mH) \* w(n) \* e^(-2 pi I / N \* kn),

Где m – номер отрезка m = 0\_M-1,

H – отступ, относительно начала предыдущего отрезка,

N – количество точек в отрезке.

Рассчитав коэффициенты для всех частот на каждом отрезке получим двумерный массив комплексных чисел. Далее в оконном преобразовании значение коэффициентов заменяют на квадрат их модуля. Следствием такой операции является двумерный массив действительных чисел представляющих влияние колебаний заданной частоты на исходную функцию на определенном отрезке времени. Визуальное изображение данного массива (обычно в виде тепловой карты) называется спектрограммой.