Проблема зашумленности аудио возникает в нашей жизни намного чаще, чем нам хотелось бы. Шумоподавление необходимо для повышения качества аудио, записанных для подкастов, музыки фильмов. Поэтому качественный аудиоредактор невозможен без опции шумоподавления. Также важной задачей является шумоподавление в реальном времени. Оно чаще всего используется в аудиоконференциях. Помимо этого, шумоподавление часто используется для предобработки звука перед применением методов автоматического распознавания речи. Тут важно понимать, что шумоподавление для улучшения качества речи для человеческого уха и шумоподавление как предобработка сигнала, для дальнейшего распознавания речи отличаются друг от друга. В последнее время, становится важным шумоподавление для улучшения качества голосовых сообщений, обмен которыми стал очень популярен в соцсетях.

Методы шумоподавления делятся на традиционные или аналитические и нейросетевые. Традиционные методы шумоподавления основываются в самом простом случае, на подавлении всех отзвуков, превышающих определенный порог громкости, либо на моделировании распределения чистой речи или шума.

Нейросетевые алгоритмы для шумоподавления делятся на две категории – на основе масок, либо генеративные. Также популярны подходы, основанные на маскировании спектрограмм. Я остановилась на последнем из этих подходов по нескольким причинам. Во-первых, несмотря на ограничения подобного подхода, таких как несовпадение частот чистого и зашумленного звука, которое не позволяет получить идеальный результат, он позволяет достигать высоких результатов в задаче шумоподавления. Также меня привлекло существование легковесных моделей для обработки спектрограмм, которые были способны к шумоподавлению. Помимо этого мне было интересно совместить обработку сигнала с обработкой изображений, к которым можно отнести спектограммы.

Отчет о проведенных экспериментах

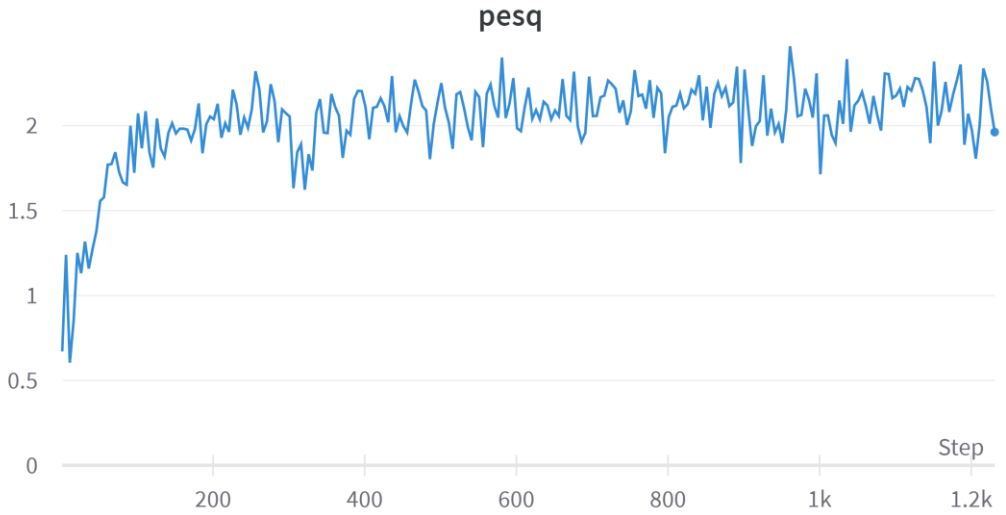
Задача шумоподавления была поставлена следующим образом: если y = s + n, где y - зашумленный аудиосигнал, подаваемый на вход модели, s - очищенный от шума сигнал, n - фоновый шум, то целевой переменной при решении задачи выступает s, представленный в виде тензора. Обучение происходило на основе двух наборов данных: commonvoice2 (содержит записи чистой человеческой речи, можно скачать по ссылке https://www.kaggle.com/datasets/danielgraham1997/commonvoice2) и urbansound8k (содержит записи с шумом, можно скачать по ссылке https://www.kaggle.com/datasets/chrisfilo/urbansound8k). Чтобы сформировать обучающий датасет случайным образом были смешаны записи чистой речи и шума, заданной длины. Каждую эпоху случайным образом формировался

новый тренировочный датасет, состоящий из 1900 зашумлённых и 1900 чистых аудиозаписей и валидационный датасет, состоящий из 100 зашумлённых и 100 чистых записей . Такая стратегия формирования данных позволила избежать переобучения модели. На этапе предобработки для получения спектрограммы использовалось кратковременное преобразование Фурье.

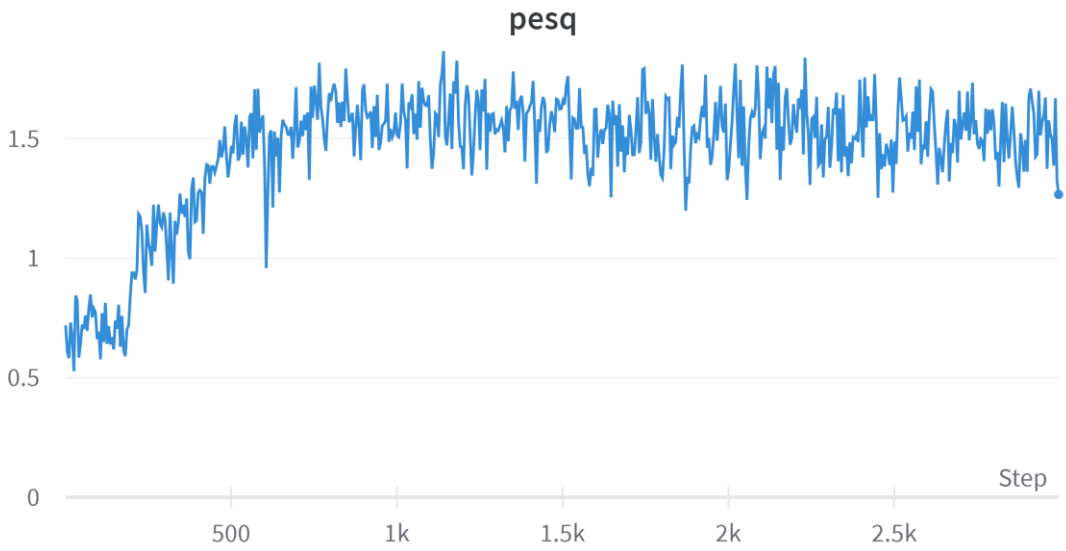
Созданные мной модели основаны на архитектуре R-CED, предложенной в статье A Fully Convolutional Neural Network for Speech Enhancement. Данная архитектура способна получать информацию из предыдущих элементов обучающей выборки во время обучения на текущем. Также в ней используются одномерные свертки. Были реализованы две модели – одна из них была максимально легковесной, а другая была остаточной нейронной сетью, с апсемплингом, реализованным при помощи бикубической интерполяции. Было решено сравнить качество работы данных моделей. Разработка и обучение моделей осуществлялись при помощи фреймворка PyTorch numpy и преобразований из библиотеки librosa. Трекинг модели проводился при помощи wandb.

В качестве функции ошибки моделей использовалась функция MAE, в качестве метрик качества – PESQ и STOI.

PESQ – метрика, используемая для оценки качества голосовой связи. Данная метрика учитывает четкость звука, фоновый шум и звуковые помехи. Принимает значения от -0.5 до 4.5, чем оно больше, тем лучше.



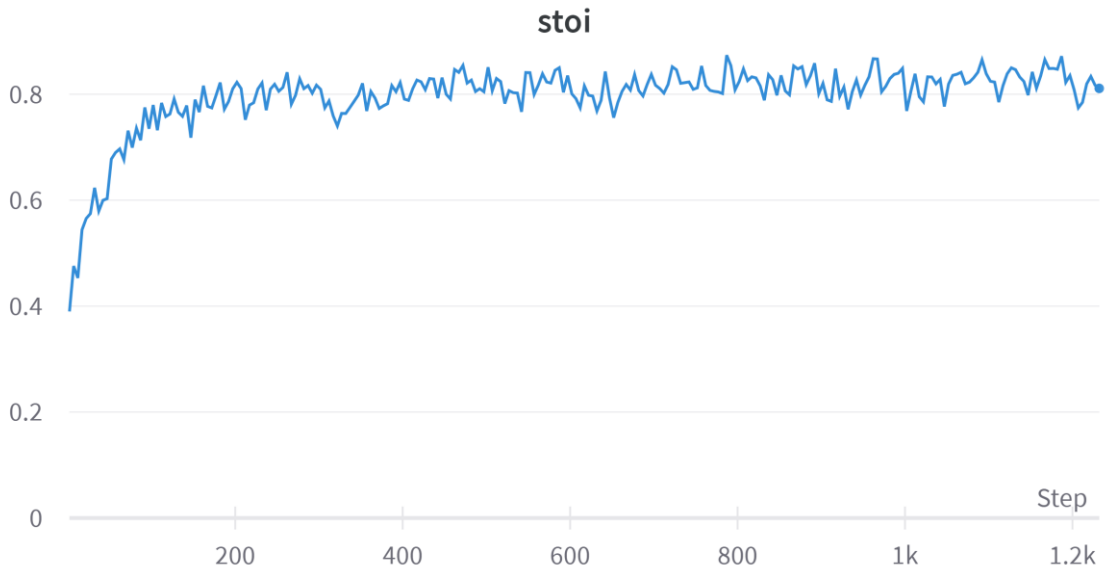
Значения метрики при обучении «тяжеловесной» модели



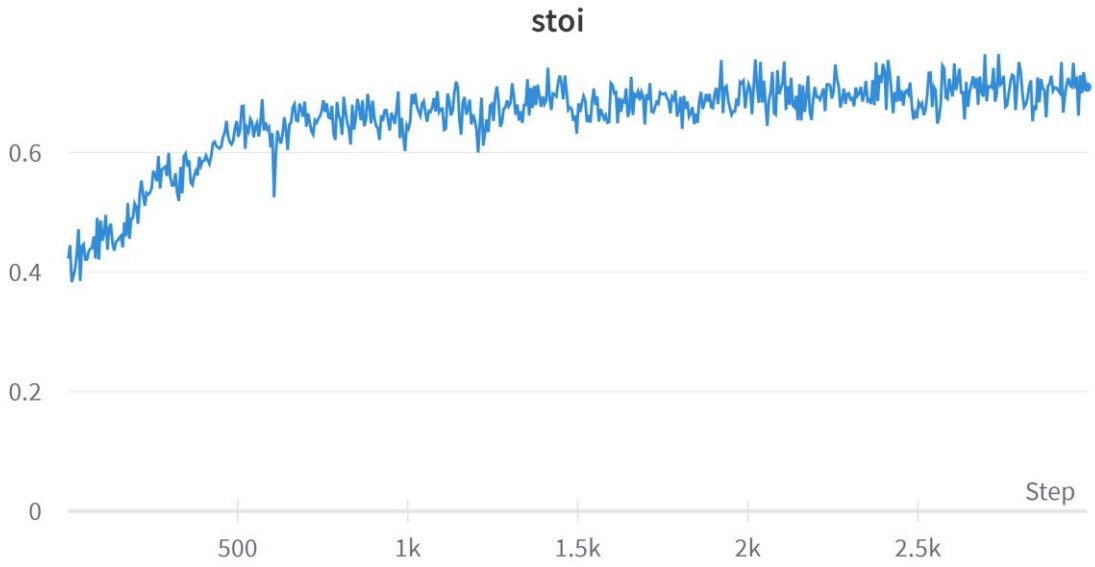
Значения метрики при обучении «легковесной» модели

Short-Time Objective Intelligibility (STOI).

Показатель разборчивости, который коррелирует с ухудшеннием речевых сигналов, например, из-за аддитивного шума или одноканального/многоканального шумоподавления. Принимает значения от 0 до 1, чем больше – тем лучше.



Значения метрики при обучении «тяжеловесной» модели



Значения метрики при обучении «легковесной» модели