|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Технологии машинного обучения \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ «Удаление шумов из аудио»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студентка ИУ5-63Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** В.В. Шаповалова

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** А.И. Канев

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

2023 г.

**Аннотация**

По ходу курса предмета «Технологии машинного обучения» было поставлено задание по созданию нейросети, способной удалить шумы из аудио. Исследование можно разделить на три части работы: разведочный анализ, предварительная обработка данных в питоне, создание алгоритма. В ходе работы будут использоваться такие библиотеки как: «TensorFlow» и «Keras». Анализ будет проводиться с помощью программы «Jupyter Python».

Будет производиться обучение модели и построены соответствующие графики и таблицы.

По итогу работы будет выполнен разведочный анализ а также создан свой алгоритм машинного обучения.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_heading=h.1fob9te)

[1. Постановка задачи 5](#_heading=h.2et92p0)

[2. Разведочный анализ 7](#_heading=h.tyjcwt)

[3.Удаление шумов из аудио 9](#_heading=h.3dy6vkm)

[ЗАКЛЮЧЕНИЕ 16](#_heading=h.u1lkzixwq4y0)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 17](#_heading=h.2s8eyo1)

# 

# **ВВЕДЕНИЕ**

Цель работы – обучение модели для удаления шумов из аудио.

Задачи:

1. Выполнить разведочный анализ, изучить актуальные алгоритмы для похожих задач;
2. Выполнить первичную обработку данных, преобразовать аудиофайлы в спектрограммы, а их в свою очередь в массивы с нормализованными значениями;
3. Обучить собственную модель для генерации чистых аудио на основе аудио с шумом в Jupyter Notebook, получить информацию о точности работы данной модели и проанализировать результаты;

# **Постановка задачи**

Чтобы поставить задачу, нужно ознакомиться с метриками, используемыми для сепарации речи. Нужные метрики зависят от конкретной задачи, но их можно разделить на 2 класса: уровня сигнала и уровня восприятия. На уровне сигнала метрики направлены на количественную оценку степени усиления сигнала или уменьшения помех. [3]

Модели улучшения речи не работают непосредственно с необработанным сигналом (т.е. сигналом во временной области), скорее они включают дискретное преобразование Фурье (DFT) в свой конвейер обработки сигналов, главным образом, в качестве первого шага для преобразования сигнала во временной области в частотную область. Эти модели признают, что речевые сигналы сильно не стационарны и их характеристики изменяются как по времени, так и по частоте. Следовательно, извлечение их частотно-временных характеристик с помощью DFT позволит лучше передать представление сигнала речевого сигнала [4].

Для отделения речи от посторонних шумов, без потери смысловой нагрузки и качества, в новых алгоритмах чаще всего используют метод градиента.

Наиболее актуальный алгоритм, представленный на рисунках 1.1 и 1.2, – преобразовать тренировочные аудиофайлы с шумами и без в waveform или спектрограммы, наложить графики друг на друга, использовать кодировщикии (encoder), декодировщики (decoder), а также сети для сепарации речи (SS Net) и усиления речи (SE Net), чтобы получить оптимальную маску шумного и чистого аудиофайлов [1].

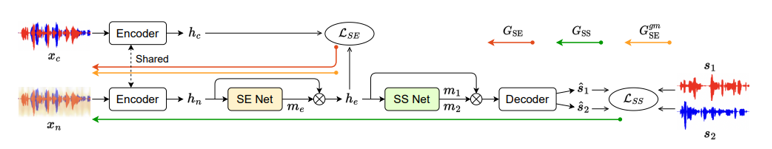


Рисунок 1.1 – Архитектура модели обучения

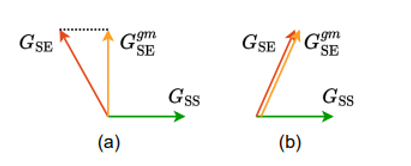


Рисунок 1.2 – Модуляция градиентов

Мне предстоит решить задачу, где есть один аудио канал, но сложнее – когда начальные данные - микс. Не так давно была создана база данных (WHAM!), в которой есть 2 канала: по одному для говорящего, а также, наложен микс звуков, которые имитируют реальную окружающую среду. Для решения задачи, где нужно отделить голоса людей и убрать шум, используют больше разновидностей исходных файлов. Естественно, в таком случае для сравнения используют более сложные функции и методы визуализации, но такой вариант наиболее приближен к реальной жизни. [2]

Обычно в таких случаях используют спектрограммы, поскольку они отражают изменение в частоте сигнала в течении времени. [5]

# **Разведочный анализ**

Поскольку в моей задаче всего один источник шума и одна дорожка голоса, я выбрала алгоритм, представленный на рисунке 2 для решения задачи.[4]

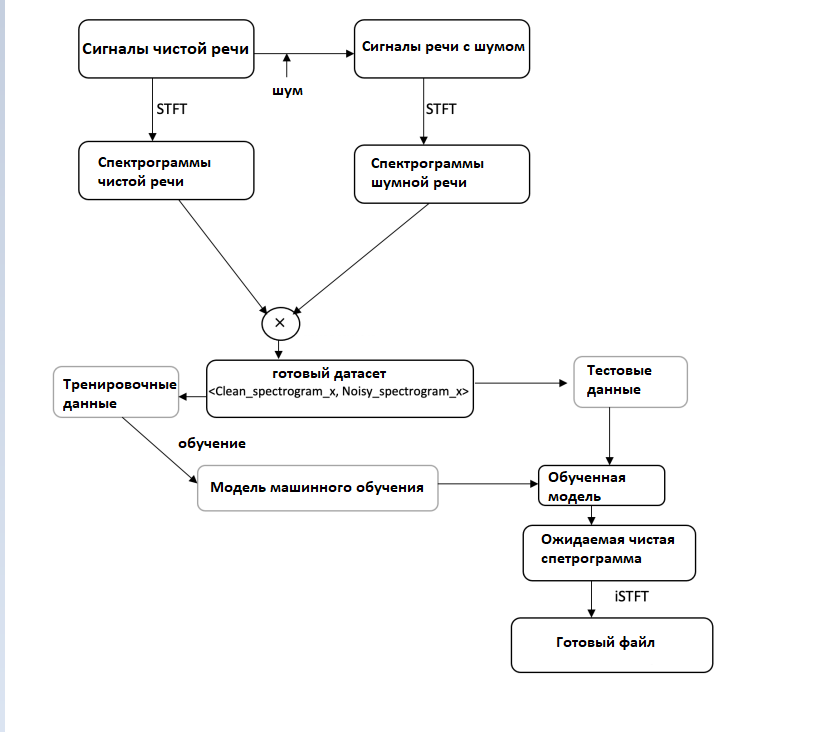


Рисунок 2 – Алгоритм решения задачи

Учитывая сигналы смешанной (зашумленной) речи, преобразуем необработанные сигналов зашумленной речи в спектрограмму. Преобразуем соответствующий сигнал чистой речи во временной области в то же представление, что и для зашумленной речи. Обучаем модель глубокого обучения, чтобы узнать, как оценивать чистые функции с учетом шумных речевых функций в качестве входных данных путем минимизации целевой функции. Учитывая новые шумные речевые характеристики, обученная модель должна оценить чистые речевые характеристики. Используя оцененные признаки чистой речи нужно восстановить ее необработанную форму сигнала, выполнив процесс, обратный процессу генерации признаков (например, используя обратное кратковременное преобразование Фурье, если признаки находятся во частотно-временной области). SFT на схеме как раз показывает, где к файлу применяется кратковременное преобразование Фурье, а iSFT – обратное преобразование.

Обучение состоит из трех этапов: кодирование, оценка маски или разделение и декодирование. Кодировщик преобразует входные смешанные сигналы во временной области в промежуточное представление, используя слои свёртки.

Оценщик масок вычисляет одну маску для говорящего. Промежуточное представление динамика получается путем умножения выходных данных кодера на его соответствующую маску. Оценщик маски состоит из блоков свертки уровня и полносвязного слоя.

Декодер преобразует промежуточные представления в разделенные во временной области речевые сигналы, используя транспонированные свёртки слои. [7].

# 

# **Удаление шумов из аудио**

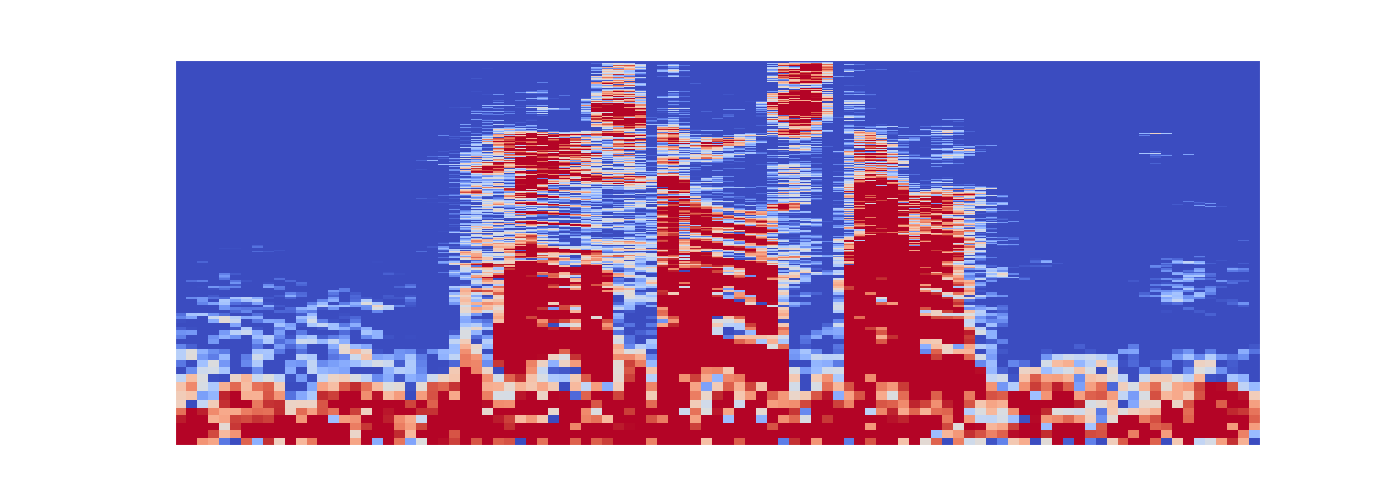
Первичная обработка данных заключалась в преобразование аудио датасета в датасет спектрограмм. Примеры спектрограмм одного и того же аудиофайла с шумом и без представлены на рисунках 3.1 и 3.2.

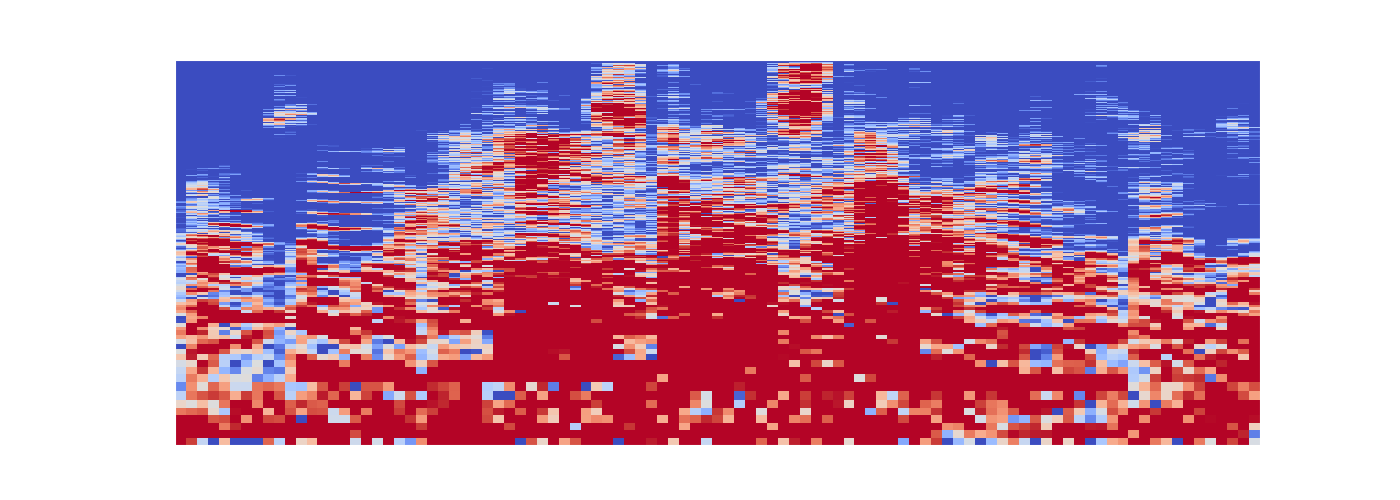
Рисунок 3.1 – Спектрограмма аудио без шума 

Рисунок 3.2 – Спектрограмма аудио с шумом

На рисунке 4 опишем функцию. которая будет открывать изображения и преобразовывать их в одинаковый размер для обучения модели, а также представлять изображение в виде массива с нормализованными значениями пискселей.

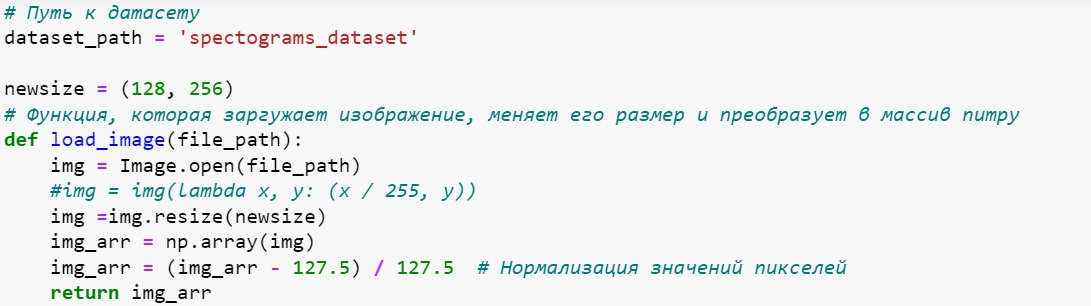


Рисунок 4 – Преобразование спектрограммы

На рисунке 5 создадим поток, который будем подавать на вход для обучения модели – спектрограммы шумных аудио файлов

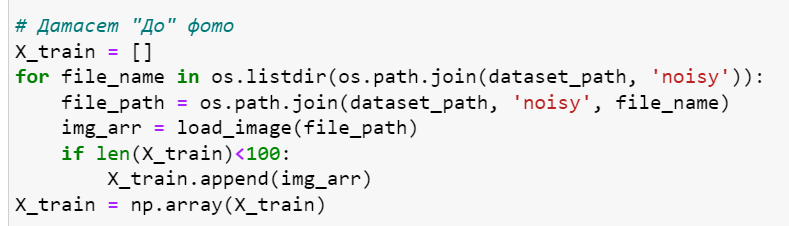


Рисунок 5 – Создание потока данных с шумом

На рисунке 6исоздадим поток, который будем подавать на выход для обучения модели –спектрограммы чистых аудио файлов

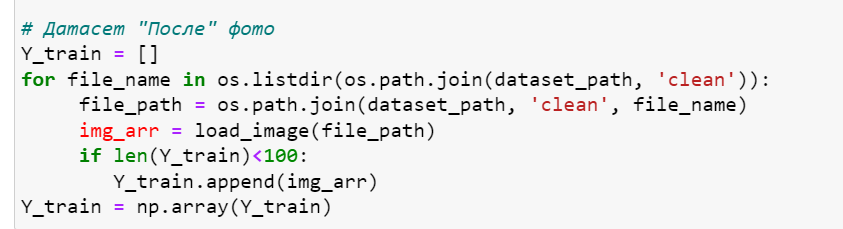


Рисунок 6 – Создание потока данных без шума

Создадим сверточную нейронную сеть – рисунок 7. В данной модели слои кодировщика и декодировщика применяют умножение матрицы свертки с входным тензором только на одной оси, вычисляя свертку вдоль оси времени. А также есть связывающий слой, где каждый нейрон связан со всеми нейронами предыдущего слоя.



Рисунок 7 – Создание нейросети

На рисунке 8 компилируем нейронную сеть c функцией потерь «средняя квадратичная ошибка», которая в данном контексте используется для измерения расхождения между исходным шумным изображением и выходным "очищенным" изображением, которое было сгенерировано автокодировщиком.

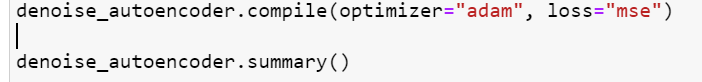


Рисунок 8 – Компиляция нейросети

На рисунке 9 выводим описание архитектуры нейронной сети, с указанием количества параметров и связей между слоями.

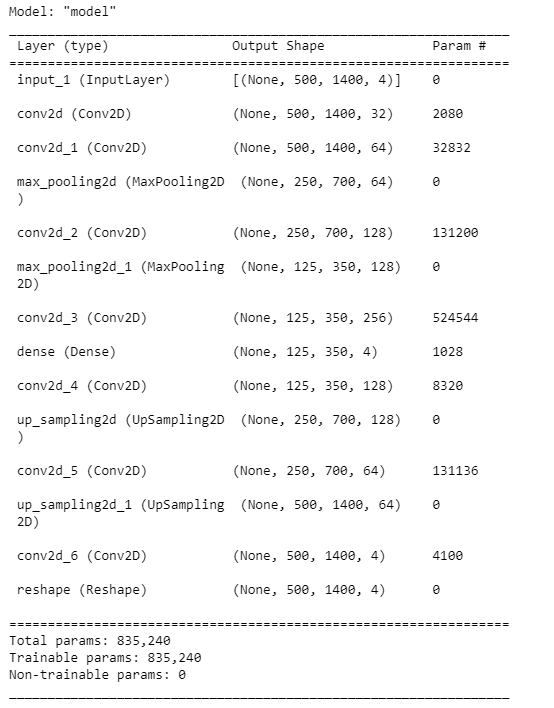


Рисунок 8 – Архитектура нейросети

На рисунке 9 определяем куда сохранять логи

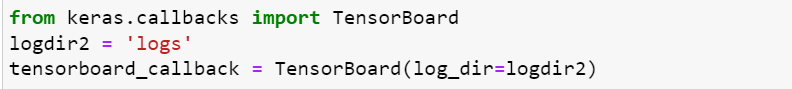


Рисунок 9 – Сохранение логов

На рисунке 10 обучаем нейросеть, разбиваем обучение на 4 эпохи, разбиваем данные на тренировочные и данные для валидации. На вход и выход подаём ранее сформированные данные.

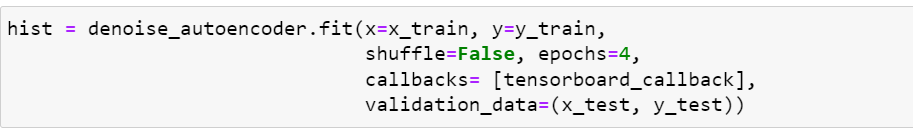


Рисунок 10 – Обучение нейросети

На рисунке 11 получаем значение метрик для каждой эпохи обучения.

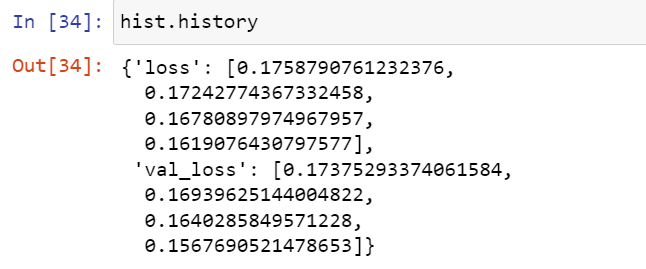


Рисунок 11 – Значение метрик

На рисунках 12.1 и 12.2. строим график функции потерь при обучении и валидации.



Рисунок 12.1 – Данные для графика

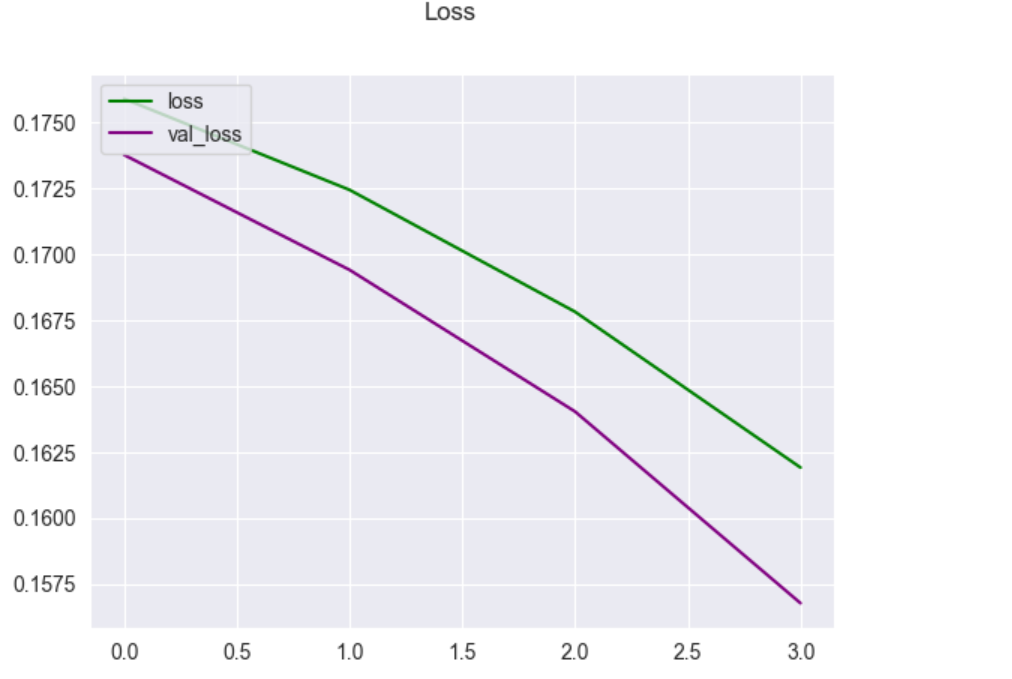


Рисунок 12.1 – График функции потерь

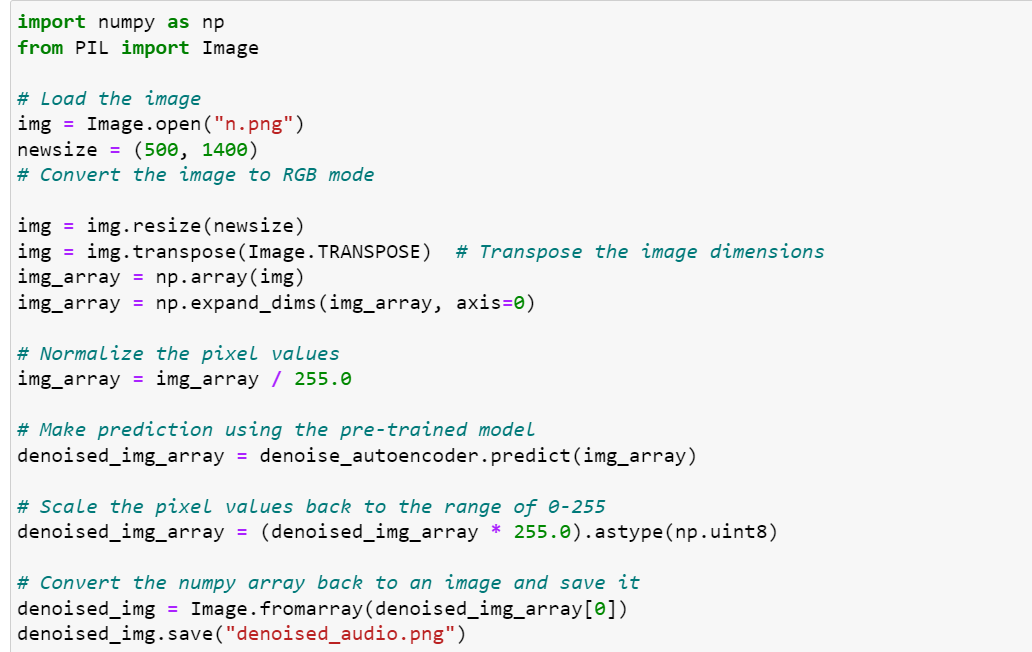
На рисунке 13 тестируем нейросеть на новой спектрограмме и сохраняем результат в формате фото. 

Рисунок 13 – Тестирование нейросети

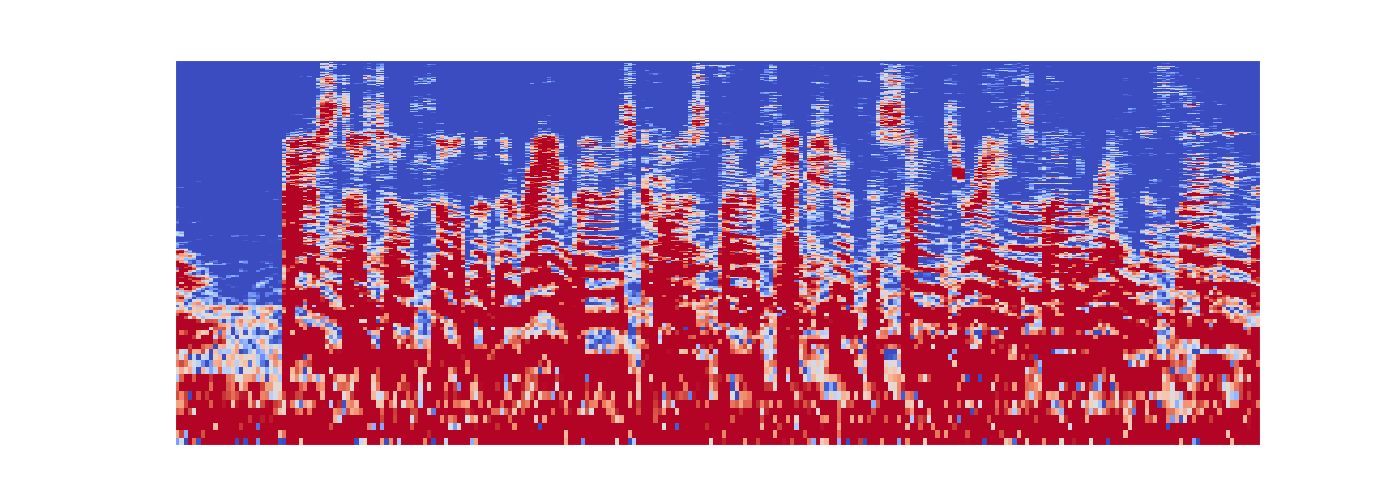
На рисунках 14.1, 14.2, 14.3 представлены результаты работы нейросети 

Рисунок 14.1 – Шумная спектрограмма на входе

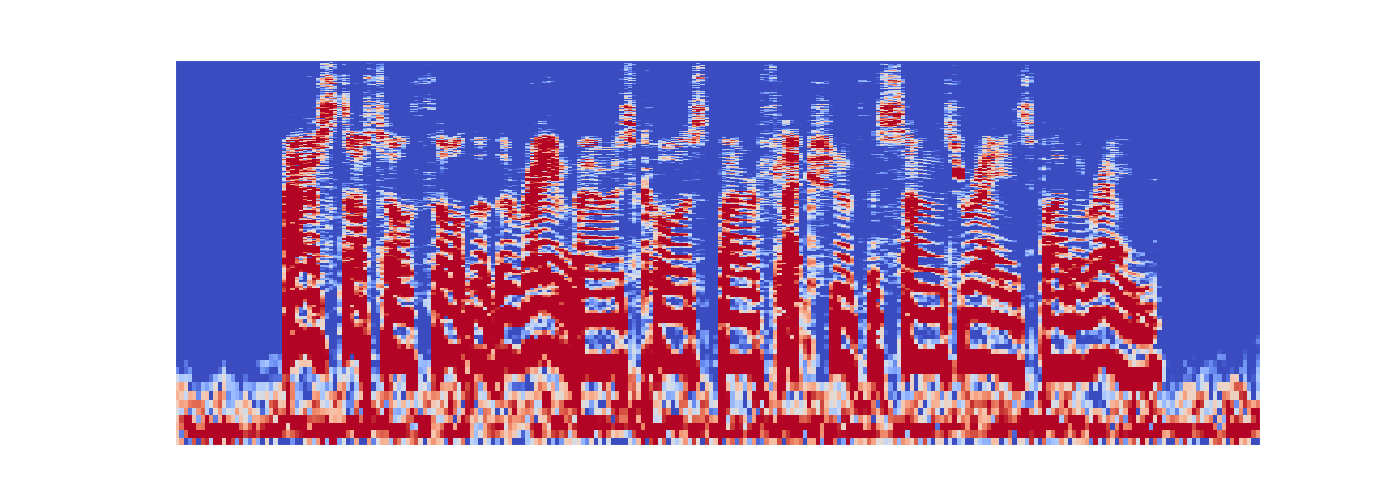


Рисунок 14.2 – Результат нейросети

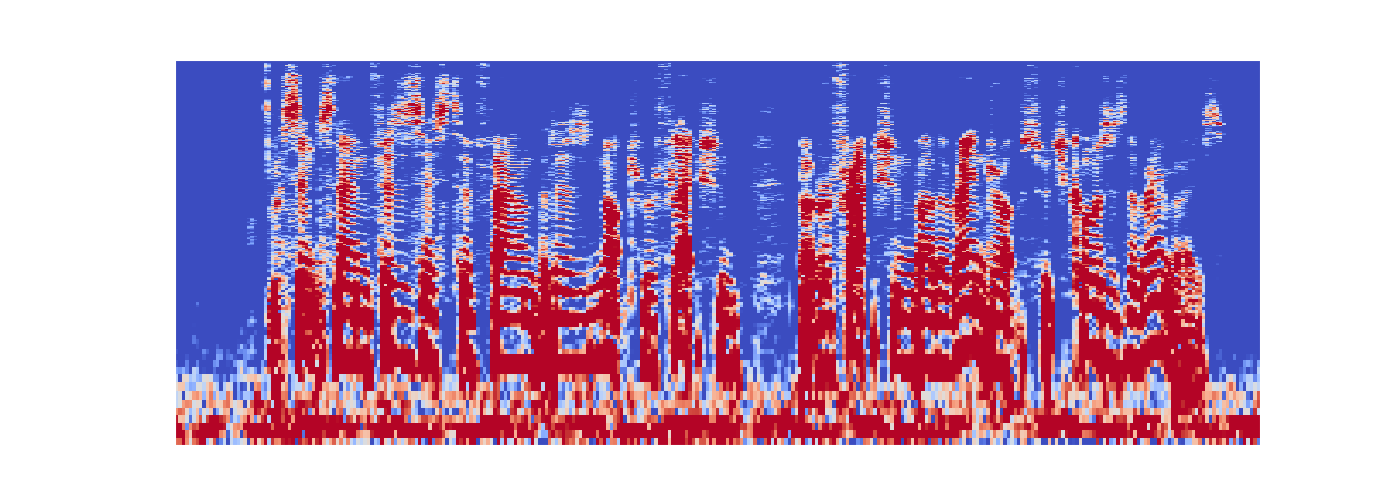


Рисунок 14.2 – Исходная спектрограмма без шума

# 

# 

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ кейса «Удаление шумов из аудиофайлов», была обучена модель генерации чистого аудио на основе аудио с шусом:

* 1. Были изучены актуальные решения данной задачи и выбран оптимальный алгоритм для данного кейса
  2. Выполнена первичная обработка данных, преобразован датасет аудио формата в данные, подходящие для обучения сверточной нейронной сети.
  3. Была обучена собственная модель для удаления шума из аудиофайлов в Jupyter Notebook, точность алгоритма не минимальная, но может быть улучшена за счёт большего количества тестовых данных.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. [Yuchen Hu](https://arxiv.org/search/eess?searchtype=author&query=Hu%2C+Y), [Chen Chen](https://arxiv.org/search/eess?searchtype=author&query=Chen%2C+C), [Heqing Zou](https://arxiv.org/search/eess?searchtype=author&query=Zou%2C+H), [Xionghu Zhong](https://arxiv.org/search/eess?searchtype=author&query=Zhong%2C+X), [Eng Siong Chng](https://arxiv.org/search/eess?searchtype=author&query=Chng%2C+E+S), «Unifying Speech Enhancement and Separation with Gradient Modulation for End-to-End Noise-Robust Speech Separation», Accepted by ICASSP 2023, <https://arxiv.org/abs/2302.11131>
2. [Gordon Wichern](https://arxiv.org/search/cs?searchtype=author&query=Wichern%2C+G), [Joe Antognini](https://arxiv.org/search/cs?searchtype=author&query=Antognini%2C+J), [Michael Flynn](https://arxiv.org/search/cs?searchtype=author&query=Flynn%2C+M), [Licheng Richard Zhu](https://arxiv.org/search/cs?searchtype=author&query=Zhu%2C+L+R), [Emmett McQuinn](https://arxiv.org/search/cs?searchtype=author&query=McQuinn%2C+E), [Dwight Crow](https://arxiv.org/search/cs?searchtype=author&query=Crow%2C+D), [Ethan Manilow](https://arxiv.org/search/cs?searchtype=author&query=Manilow%2C+E), [Jonathan Le Roux](https://arxiv.org/search/cs?searchtype=author&query=Roux%2C+J+L) «WHAM!: Extending Speech Separation to Noisy Environments»,   
   Accepted for publication at Interspeech 2019, https://arxiv.org/abs/1907.01160
3. [DeLiang Wang](https://arxiv.org/search/cs?searchtype=author&query=Wang%2C+D), [Jitong Chen](https://arxiv.org/search/cs?searchtype=author&query=Chen%2C+J), «Supervised Speech Separation Based on Deep Learning: An Overview», 15 Jun 2018, <https://arxiv.org/abs/1708.07524>
4. [Peter Ochieng](https://arxiv.org/search/cs?searchtype=author&query=Ochieng%2C+P), «Deep neural network techniques for monaural speech enhancement: state of the art analysis», 1 Dec 2022 <https://arxiv.org/abs/2212.00369>
5. [Yi Luo](https://arxiv.org/search/cs?searchtype=author&query=Luo%2C+Y), [Nima Mesgarani](https://arxiv.org/search/cs?searchtype=author&query=Mesgarani%2C+N), «Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation», Accepted by IEEE/ACM Transactions on Audio, Speech and Language Processing 2019, <https://arxiv.org/abs/1809.07454>
6. Ruohan Gao1,2 Kristen Grauman1,3 1The University of Texas at Austin 2Stanford University 3Facebook AI Research, « VISUALVOICE: Audio-Visual Speech Separation with Cross-Modal Consistency», In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021
7. MathWorks, End-to-End Deep Speech Separation, https://www.mathworks.com/help/deeplearning/ug/end-to-end-deep-speech-separation.html