УДК: 534.4
OECD: 1.03

Оптимизация параметров предобработки аудиосигналов для нейросетевой классификации шумов

Майер Д.В.[1*], Стаценко Л.Г.[2]
[1]Аспирант, [2]Д.ф.-м.н., профессор
[1,2]Департамента электроники, телекоммуникации и приборостроения, Политехнический институт, Дальневосточный Федеральный университет, г. Владивосток, РФ

Аннотация

В работе представлен анализ влияния параметров предобработки аудиосигналов на эффективность классификации шумов с использованием сверточно-рекуррентной нейронной сети. В качестве исследуемых характеристик рассматривались ширина временного окна и различные виды представления спектрограмм с использованием быстрого преобразования Фурье, а также Мел- и Барк-шкал. На основе созданного датасета, включающего 10 классов городских шумов, проведена серия экспериментов с варьированием ширины временных кадров (0,1–0,3 секунды) и формата спектрального представления. Оценка качества осуществлялась с использованием метрик Accuracy, Precision, Recall и F1-score. Результаты показали, что наилучшие показатели продемонстрировала модель на основе Мел-спектрограмм при ширине временного кадра 0,2 секунды, обеспечившая оптимальный баланс между временной и частотной разрешающей способностью. Сделан вывод о влиянии параметров спектрального анализа на качество работы нейронных сетей.

Ключевые слова: городской шум, спектрограмма, сверточно-рекуррентная нейронная сеть, классификация шумов, предобработка аудиосигналов, машинное обучение

Optimization of audio signal preprocessing parameters for neural network noise classification

Maier D.V.[1], Statsenko L.G.[2]*
[1]Asparant, [2]Ph.D., Full Professor
[1,2]Department of Electronics, Telecommunications and Instrumentation, Polytechnic Institute, Far Eastern Federal University, Vladivostok, Russia

Abstract

The paper presents an analysis of the influence of audio signal preprocessing parameters on the effectiveness of noise classification using a convolutional recurrent neural network. The investigated characteristics were the width of the time window and various methods for constructing spectrograms: short-time Fourier transform, Mel and Bark spectrograms. Based on the created dataset, which includes 10 classes of urban noises, a series of experiments were conducted with varying the duration of time frames (0.1–0.3 seconds) and the spectral representation format. The quality assessment was carried out using Accuracy, Precision, Recall and F1-score metrics. The results showed that the best performance was demonstrated by a model based on Mel spectrograms with a frame length of 0.2 seconds, which provided

**E-mail: di.7.3@yandex.ru (Майер Д.В.)*

an optimal balance between time and frequency resolution. The conclusion is made about the influence of spectral analysis parameters on the quality of neural networks.

Keywords: *urban noise, spectrogram, convolutional recurrent neural network, noise classification, audio signal preprocessing, machine learning*

Введение

Вопрос влияния шума на человека до сих пор недостаточно изучен, несмотря на явную необходимость контроля за постоянно растущим уровнем акустического загрязнения. Для обеспечения комфортных условий жизни населения и сохранения здоровья окружающей среды становится всё более важной задача не только контроля шума, но и разработки эффективных мер его снижения. Первым и основным шагом борьбы с акустическим загрязнением является детектирование шума, определение его источников и их последующая классификация, поэтому данное исследование нацелено на улучшение методов контроля уровня городских шумов, что в последствие может быть использовано для разработки стратегий и мероприятий для улучшения качества жизни и снижения уровня акустического загрязнения [1-3].

В качестве инструмента анализа шумовой среды предлагается использование методов машинного обучения и искусственного интеллекта. Такие подходы позволяют эффективно обрабатывать большие объёмы данных, строить сложные модели прогнозирования и классификации, а также учитывать широкий спектр факторов, влияющих на восприятие шума человеком. Современные методы обработки аудиосигналов широко используют нейронные сети, эффективность которых во многом зависит от качества предварительной обработки данных. Одной из ключевых задач является выбор оптимального способа представления аудиосигнала в спектральной форме, если обработка информации осуществляется в виде изображения.

Особое значение данная проблема имеет при анализе шумов. Шумовые сигналы отличаются высокой вариативностью, сложностью спектрального состава и отсутствием ярко выраженной структуры, характерной для речи или музыкальных сигналов. В связи с этим выбор адекватного способа их спектрального представления и оптимизация параметров предобработки становятся критически важными задачами для обеспечения корректной работы нейронных сетей. Таким образом, исследование параметров спектрального анализа шумов и их влияния на эффективность нейронных сетей представляет собой актуальную научную и практическую задачу.

Цель работы заключается в исследовании влияния ширины временного окна и различных методов построения спектрограмм на эффективность акустического представления шумовых сигналов с дальнейшим их сравнительным анализом. Результатом работы является определение оптимальных параметров и подходов для предобработки аудиоданных, используемых в нейронных сетях. Эффективность оценивается на основе качества работы нейронных сетей при обработке и распознавании шумов.

Искусственный интеллект активно применяется для анализа и управления шумом, позволяя классифицировать его источники по спектральным характеристикам и использовать полученные данные для мониторинга городской среды [4]. Несмотря на высокий потенциал, методы машинного обучения пока недостаточно широко применяются в области оценки шумов, хотя именно глубокие нейронные сети уже доказали эффективность в задачах классификации звуков и требуют адаптации

к реальным условиям.

Одним из самых распространенных методов обработки звуков и шума сейчас является анализ спектрограмм аудиозаписей как изображений [5-6]. Часто и достаточно эффективно применяются глубокие сверточные нейронные сети (CNN) для анализа окружающего звукового фона на основе спектрограмм записей. Модель глубокого обучения способна с высокой точностью классифицировать типы звуковой среды, различая, например, транспортный шум, голосовые разговоры, тишину или шум толпы [7]. CNN отвечает за извлечение локальных признаков со входной спектрограммы. Именно сверточные слои обнаруживают характерные паттерны, уникальные для разных типов шумов. Такие признаки могут быть невидимы при классическом ручном анализе, но хорошо выделяются сетью за счёт автоматического обучения фильтров.

Однако звуковые сигналы имеют сложную временно-частотную структуру, которую необходимо анализировать не только как статичное изображение (например, спектрограмму), но и учитывать последовательности изменений во времени. Важно выделять не только отдельные частотные компоненты, но и характер их изменения, длительность, повторяемость – всё то, что определяет суть звукового события. Для решения этой задачи современные подходы используют рекуррентные нейронные сети (RNN), способные учитывать временные закономерности, потому что любой звук – это не только совокупность частот, но и их динамика во времени. Рекуррентные сети позволяют учитывать контекст, что особенно важно для различения похожих по частотным характеристикам, но разных по смыслу шумов.

Высокие результаты показывают смешанные модели – сверточно-рекуррентные нейронные сети (CRNN) [8-9]. Авторы данных работ предлагают модель, которая способна определять не только какие звуковые события происходят, но и откуда именно исходит звук. Особое внимание уделено способности модели обрабатывать перекрывающиеся звуковые источники, что ранее было одной из сложнейших задач в акустическом анализе.

1 Методы и инструменты исследования

Учитывая популярность и эффективность сверточно-рекуррентной нейронной сети для задачи обнаружения и классификации шумов в городской среде, в данной работе предлагается использование именно CRNN, которая является обоснованным выбором, поскольку именно такая архитектура максимально учитывает специфику акустических данных. Ещё одним преимуществом выбранной архитектуры является её универсальность и масштабируемость. При необходимости модель можно дополнительно усложнить, добавляя слои или используя более современные решения.

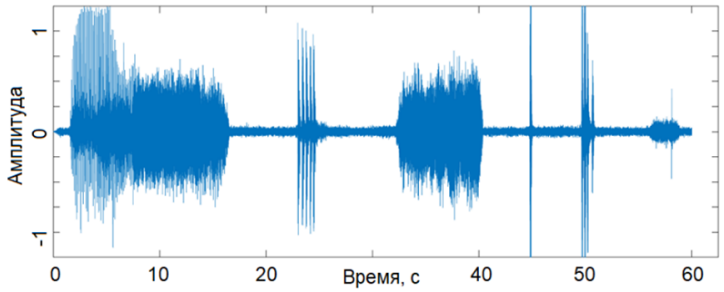
В качестве входных «сырых» данных предлагается использование минутных аудиодорожек, которые содержат в себе временные отрезки шумов определённых классов. Таким образом, эти записи представляют собой набор различных звуковых событий. Частота дискретизации (Sample Rate) каждого аудиофайла составляет 32 кГц, что необходимо для более точного анализа сложных акустических сигналов, включая высокочастотные шумы, например, сирены.

К каждому аудиофайлу привязан отдельный CSV-файл с аннотациями, содержащий временные метки звуковых событий. В этих CSV-файлах каждая строка соответствует отдельному звуковому событию и содержит следующую информацию: время начала события (Start), время окончания события (End), класс звукового события (например, «Гудок», «Лай собак», «Речь» и др.). Таким образом, CSV-файл выступает в роли точной расшифровки содержимого аудиофайла, позволяя знать, в какой момент времени и какой именно звук присутствует в записи.

Такое представление данных обеспечивает точное соответствие между аудиосигналом и его содержанием, необходимое для обучения модели. Во время подготовки обучающего набора спектрограмма аудиофайла делится на временные кадры, и каждому кадру сопоставляется метка класса на основе информации из CSV-аннотации. Если в данный момент времени в записи отсутствует целевое звуковое событие, кадру присваивается метка «тишина».

Использование именно такого подхода – аудиофайлов в связке с CSV-метками – является стандартом в задачах классификации звуковых событий (Sound Event Detection – SED), поскольку позволяет обеспечить строгую синхронизацию данных и меток.

Специально для данного исследования был создан уникальный датасет из 63 минутных аудиодорожек в формате .wav, при использовании разных аудиоданных из открытых источников. В качестве классов было принято решение использовать 10 разных видов акустических событий: вертолет, выстрел, гудок, дождь, лай собак, мотоцикл, плач ребенка, поезд, речь и сирена. Для наглядности предлагается рассмотреть один сигнал из датасета – сигнал №1, который показан на рисунке 1.



	File	Start	End	Class
1	1	1.583	16.519	Вертолет
2	1	22.653	25.67	Выстрел
3	1	31.793	40.381	Мотоцикл
4	1	44.613	45.317	Лай_собак
5	1	49.467	51.055	Лай_собак
6	1	55.941	58.941	Дождь

Временные метки

Рисунок 1 – Осциллограмма сигнала №1 из датасета и временные метки расположения классов сигналов для данного сигнала

Так как аудиофайл имеет разное качественное наполнение, то невозможно оценить и классифицировать весь аудиосигнал единой меткой. Чтобы сформировать целевые значения и обучить сеть, необходимо сопоставить временные отметки с кадрами. Общая продолжительность каждого файла составляет 60 секунд; пусть файл будет разделён на 600 кадров для формирования целевых меток, то есть модель будет делать предсказание каждые 0,1 секунды. Каждый временной фрейм соответствует одному значению метки (классу или «тишина»).

Однако такие данные невозможно использовать для заданной модели без предварительной обработки, поэтому извлечение признаков из аудиосигнала предлагается осуществлять с использованием преобразования Фурье, а именно через функцию быстрого преобразования Фурье (БПФ, англ. STFT – short-time Fourier transform). Преобразование Фурье имеет следующие характеристики: длину окна 512 точек и шаг 400 точек, при использовании оконной функции Хэмминга.

Результатом преобразования становится набор частотных ячеек (или бинов). Каждый бин представляет собой определённый диапазон частот и показывает, насколько сильно этот диапазон представлен в сигнале. Одна частотная ячейка (бин) представляет собой индекс элемента спектра, за которым закреплена полоса частот фиксированной ширины. Чем выше номер бина, тем выше частота, которую он описывает. В данном случае частота дискретизации сигнала составляет 32 кГц, а длина окна БПФ – 512, поэтому то ширина одного бина около 62,5 Гц.

После получения спектра производится преобразование амплитудных значений

в логарифмическую шкалу. Это преобразование необходимо, так как человеческое ухо воспринимает звук не линейно, а логарифмически, и такая шкала более точно отражает субъективное восприятие громкости. Логарифмирование также стабилизирует данные, уменьшая диапазон значений и делая их более подходящими для подачи на вход нейронной сети. Спектрограмма сигнала №1 показана на рисунке 2.

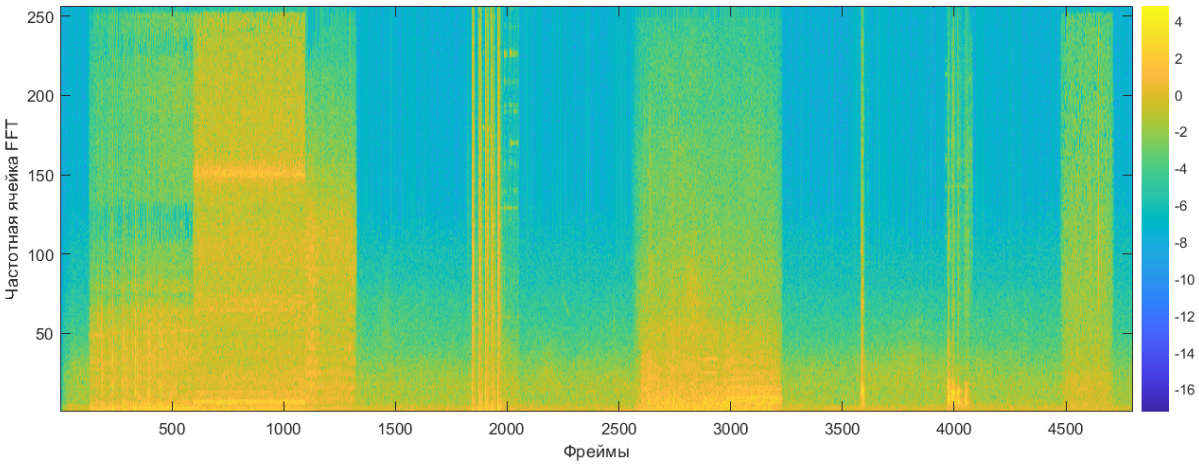


Рисунок 2 – Спектрограмма сигнала №1 из датасета

В качестве альтернативы предлагается использовать мел-шкалу и Барк-шкалу, получая мел-спектрограмму и Барк-спектрограмму соответственно. Данные формы отображения лучше согласуются с восприятием звука человеком. Обе шкалы являются психоакустическими, то есть моделируют восприятие частоты человеческим ухом. На мел-шкале низкие частоты представлены более детально, а высокие – более сжаты. Барк-спектрограмма основана на шкале Барк, которая аналогично связана с физиологией слуха человека и с критическими полосами слухового восприятия. Ухо делит звуковой спектр на критические полосы, в пределах которых звуки воспринимаются как взаимодействующие между собой.

Для корректной подачи данных на вход модели все спектрограммы масштабируются до размера 256×4800. Мел- и Барк-спектрограммы сигнала №1 показаны на рисунке 3.

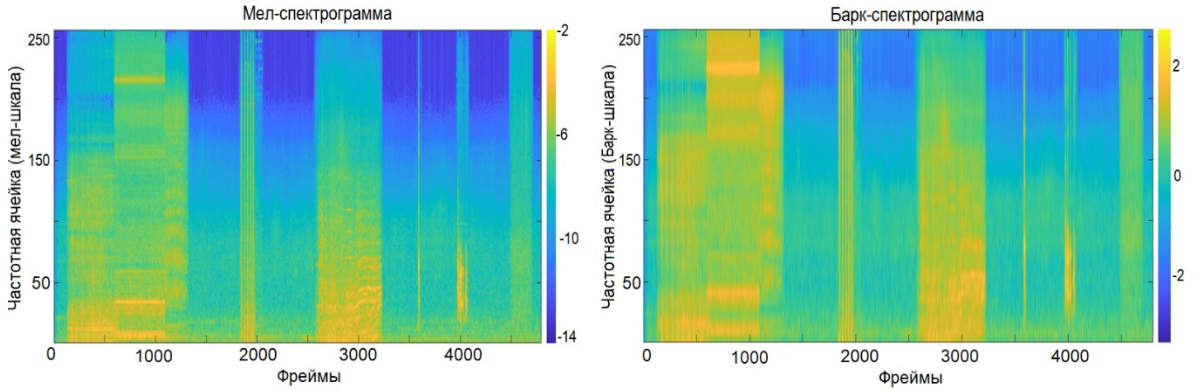


Рисунок 3 – Мел-спектрограмма и Барк-спектрограмма сигнала №1 из датасета

Результирующие спектрограммы имеют двумерную форму: временные кадры и частотные компоненты. Эта структура особенно хорошо подходит для обработки с помощью сверточных нейронных сетей, так как спектрограмма по сути является

изображением, а CNN изначально создавались для выделения признаков на изображениях. Такая схема обеспечивает максимальное сохранение информации как о частотном составе, так и о временной структуре звука, что критически важно для задач точной классификации и последующего анализа акустических характеристик городского шума.

Обработка данных и построение модели осуществляется на базе прикладной программы MATLAB, которая хорошо подходит для обработки сигналов, аудио, визуализации и реализации алгоритмов машинного обучения. В данном разделе будет показана структура нейронной сети.

Вся сеть состоит из двух последовательных частей: сверточной сети и реккурентной сети. На вход CNN подаются спектрограммы аудиосигналов, представленные как двумерные изображения. Сеть состоит из четырёх сверточных блоков, постепенно увеличивающих число фильтров от 64 до 512 и выделяющих всё более сложные признаки звукового сигнала. Первые блоки активно сжимают данные по частоте и времени, чтобы выявить характерные временно-частотные паттерны, а последние блоки увеличивают глубину признаков, формируя компактное и абстрактное представление сигнала без изменения пространственной размерности.

После сверточных блоков применяется операция изменения формы (reshape), преобразующая выход CNN в последовательность, подходящую для подачи на рекуррентные слои.

Следующая часть модели относится к классу рекуррентных нейронных сетей и построена на базе трёхслойной структуры с использованием двунаправленных рекуррентных слоёв на основе элементов GRU (Gated Recurrent Units). Модель предназначена для обработки последовательных данных, поступающих после извлечения признаков сверточной частью сети, и обеспечивает детектирование и классификацию звуковых событий во временной последовательности. Она включает три двунаправленных слоя GRU для учёта контекста сигнала в обоих временных направлениях, за которыми следуют три полносвязных слоя для обобщения признаков. Архитектура завершается выходным слоем, соответствующим числу классов звуковых событий.

Основной цикл обучения нейронной сети реализован итеративно с контролем качества на валидации, обновлением параметров методом Adam и сохранением лучших результатов. Обучение продолжается до достижения 200 эпох или остановки улучшений на валидационном наборе в течение 10 эпох. По окончании каждой эпохи модель оценивается, и при снижении ошибки сохраняется как лучший вариант; при отсутствии прогресса увеличивается счётчик «плохих эпох», что обеспечивает раннюю остановку.

Стоит указать, что несмотря на наличие 10 классов, присутствует 11-ый класс – «Тишина». Он не фигурирует в классификации, так как он определяется механически с помощью порогового значения вероятности, а не как отдельный класс в выходном слое сети. Модель на выходе формирует вероятности принадлежности каждого временного кадра к одному из 10 звуковых классов, после получения предсказаний выполняется постобработка, где все вероятности ниже порога (по умолчанию 0,3) обнуляются, то есть считаются незначимыми. Таким образом считается, что в эти моменты звуковых событий нет, поэтому в такие моменты происходит «тишина». Такой способ позволяет модели быть гибкой и не вынуждает ее всегда выбирать один из классов.

Итогом работы является сохранённый файл с весами обученной нейросети, который содержит все параметры сети, полученные в процессе обучения. Для анализа качества работы одной и той же модели будут изменятся такие показатели предобработки как вид спектрограммы и количество временных кадров. Результаты обучения позволят определить оптимальную ширину кадра для классификации шумов и наиболее успешный вид спектрограммы.

2 Результаты обучения

Для первого обучения выбраны два отличительных параметра обработки сигнала: использование логарифмической спектрограммы в результате использования БПФ (STFT) и разделение сигнала на 600 временных кадров. Такую сборку модели можно обозначить как STFT/600.

После успешного обучения модель оценивается с помощью специальных метрик, наиболее наглядной из которых является матрица ошибок (confusion matrix). Она представляет собой таблицу, где строки соответствуют истинным классам, а столбцы – предсказанным. Значения на диагонали отражают число верных классификаций, вне диагонали – ошибки, что позволяет выявить хорошо распознаваемые классы и классы, чаще всего путаемые моделью.

На рисунке 4 показана нормированная по строкам матрица ошибок для модели STFT/600. Можно заметить, что достаточно много модель сделала попаданий, однако, если обратить внимание на класс «Гудок», то можно увидеть, что в этой строке много значений вне диагонали, значит модель часто путает гудки с другими звуками. Можно сделать вывод, что данная модель абсолютно не может распознать сигнал гудка машины, путая его с шумом вертолета и мотоцикла.

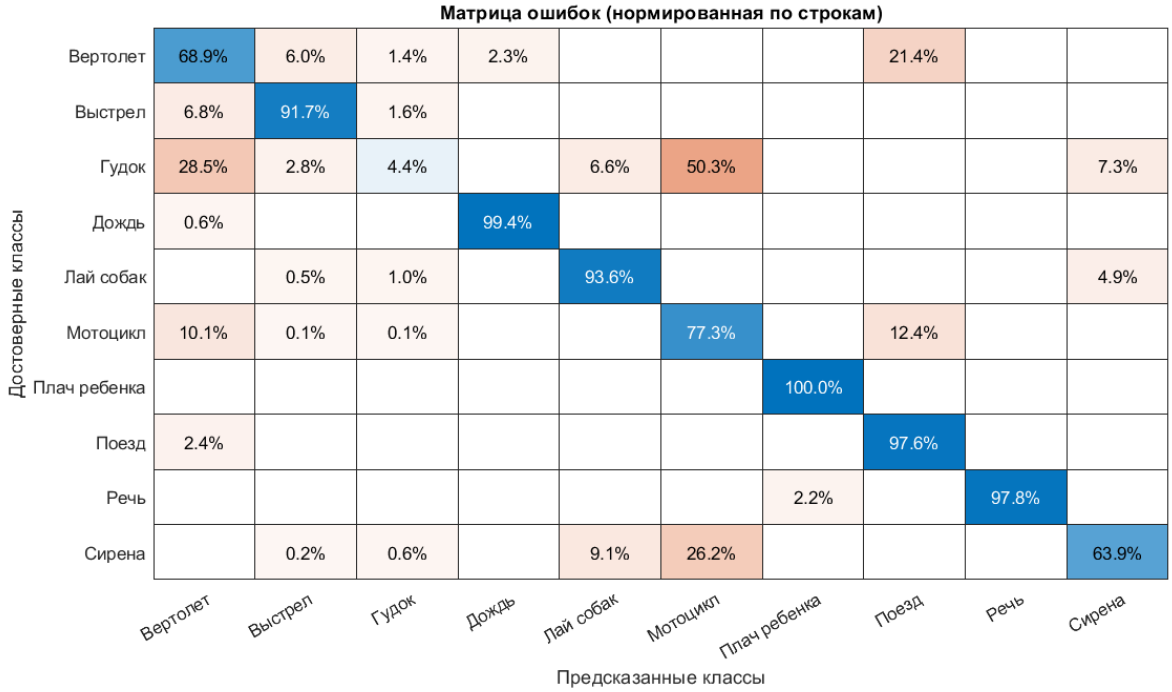


Рисунок 4 – Нормированная матрица ошибок для модели STFT/600

На основе матрицы ошибок рассчитываются обобщающие метрики качества модели. Базовой является точность (Accuracy), показывающая долю верных предсказаний, однако при несбалансированных классах она недостаточно информативна. В таких случаях применяются Precision и Recall: первая характеризует долю корректных предсказаний для класса, вторая – полноту выявления истинных событий. Для комплексной оценки используется F1-score, представляющий гармоническое среднее Precision и Recall, что особенно важно при дисбалансе классов.

В случае многоклассовой классификации метрики Precision, Recall и F1 можно вычислить только для каждого класса отдельно, что показано на рисунке 5.

В целом модель при данных характеристиках входных данных показывает хорошие

результаты для большинства классов, но есть несколько классов с заметно худшими показателями. Сначала стоит отметить, что модель почти идеально распознает дождь, плач ребенка и речь. Проблемным классом является шум гудка машины – модель практически не распознаёт гудки, путает их с другими шумами.

Class	Precision	Recall	F1
"Вертолет"	0.68005	0.68928	0.68464
"Выстрел"	0.89394	0.9165	0.90508
"Гудок"	0.35897	0.044304	0.078873
"Дождь"	0.98372	0.99419	0.98893
"Лай собак"	0.70632	0.93596	0.80508
"Мотоцикл"	0.69722	0.7731	0.7332
"Плач ребенка"	0.97535	1	0.98752
"Поезд"	0.59211	0.9759	0.73703
"Речь"	1	0.97814	0.98895
"Сирена"	0.92534	0.63906	0.75601

Рисунок 5 – Метрики по классам для модели STFT/600

Чтобы оценить модель в целом, можно использовать Accuracy и дополнительные метрики, такие как Macro Average (среднее арифметическое F1-score по всем классам), Weighted Average (как Macro, но учитывает, сколько примеров в каждом классе (важно при дисбалансе классов)).

Общие результаты для модели STFT/600:

– Accuracy: 82,72 %;

– Macro F1: 76,65 %;

– Weighted F1: 81,21 %.

Можно сделать вывод, что модель хорошо работает в среднем (Accuracy $>$ 80 %), особенно с основными, наиболее частыми классами. Присутствует проблема с отдельными классами, о чем говорит низкий показатель Macro F1. Для практического применения в задачах акустического мониторинга модель уже пригодна, но требует доработки по слабораспознаваемым классам.

Далее будут рассмотрены другие вариации признаков в более сокращенном формате. Для повышения качества работы модели принято решение уменьшить частоту предсказаний модели в 2 раза, то есть теперь модель будет делать предсказание каждые 0,2 секунды, а каждый файл будет разделён на 300 кадров. Такую сборку модели можно обозначить как STFT/300. Такое изменение может уменьшить количество одиночных всплесков ложных срабатываний и увеличить вероятность корректного распознавания шума.

На рисунке 6 показана нормированная матрица ошибок для модели STFT/300, где можно заметить улучшение качества распознавания сигнала «Гудок». Оно недостаточно высокое, чтобы быть удовлетворительным, но уже значительно выше предыдущего результата. Аналогично вырос уровень точности предсказывания других классов, так в определении «Речи» модель не сделала ни одного промаха в определении истинного класса, соответственно Recall («Речь») равен 100 %.

Общие результаты для модели STFT/300:

– Accuracy: 87,65 %;

– Macro F1: 84,05 %;
– Weighted F1: 87,08 %.

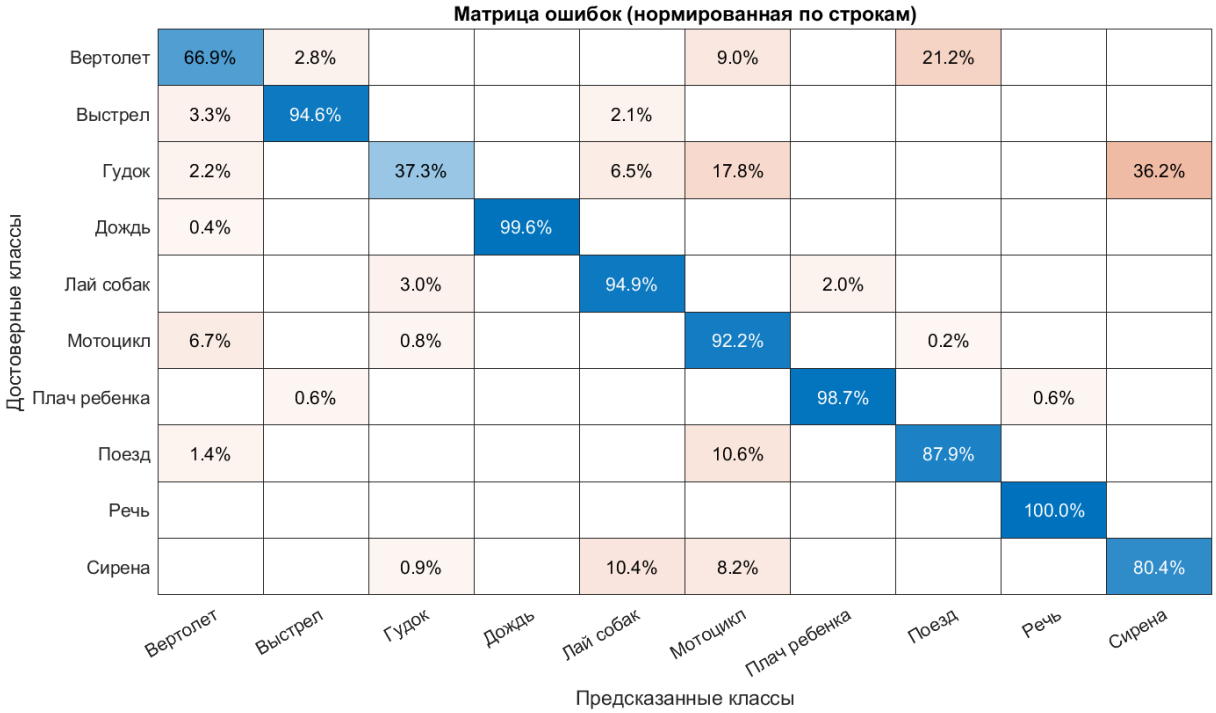


Рисунок 6 – Нормированная матрица ошибок для модели STFT/300

Чтобы определить оптимальный временной интервал распознавания, необходимо еще уменьшить количество временных кадров до 200 – теперь модель будет делать предсказание каждые 0,3 секунды, а каждый файл будет разделён на 200 кадров. Такую сборку модели можно обозначить как STFT/200, нормированная матрица ошибок для которой показана на рисунке 7, где можно заметить абсолютное игнорирование моделью сигнала «Лай собак». Данный факт является критической ошибкой, так как модель совсем не распознаёт этот класс. Модель полностью спутала данный класс с «Сиреной». Можно предположить, что 200 кадров – слишком мало, чтобы адекватно уловить данный класс, так как один «гав» собаки длится в среднем 0,2 секунды. Возможно разбиение кадров не совпало с границами лая, и поэтому модель совсем не научилась его распознавать.

Общие результаты для модели STFT/200:
– Accuracy: 76,65 %;
– Macro F1: 68,29 %;
– Weighted F1: 75,43 %.

Видно, что общие показатели упали более, чем на 10% каждый по сравнению с результатами предыдущей модели. Можно утверждать, что наиболее оптимальное количество временных кадров для сигнала с набором шумов длительностью 1 минут соответствует 300. Для повышения качества работы модели можно изменить не только размер данных, но и внутреннее наполнение. Далее будут проанализированы результаты обучения модели на основе мел- и Барк-спектрограмм. Сначала предлагается рассмотреть итоги работы модели, обучаемой на основе мел-спектрограмм. Такую сборку модели можно обозначить как Mel/300.

На рисунке 8 показана нормированная матрица ошибок для модели Mel/300. В данной таблице можно заметить улучшение качества распознавания сигналов «Гудок»

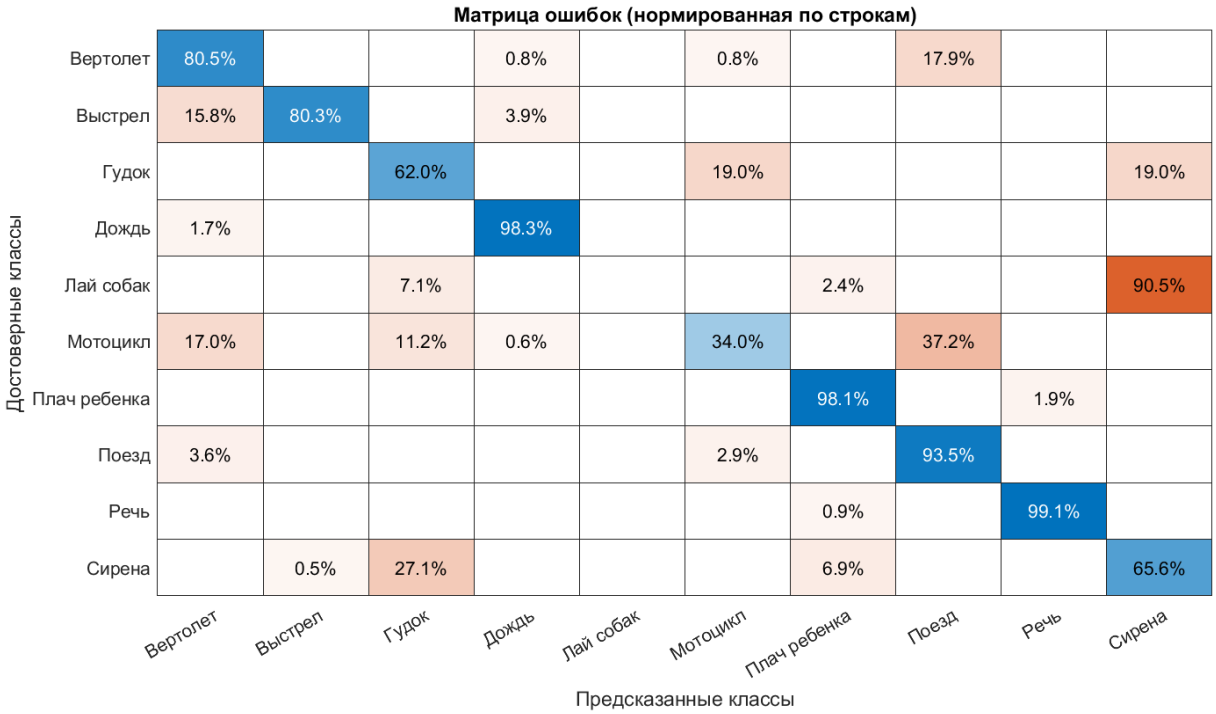


Рисунок 7 – Нормированная матрица ошибок для модели STFT/200

и «Лай собак». Точность определения каждого класса теперь составляет не менее 72%, что пока является самым высоким результатом из всех моделей.

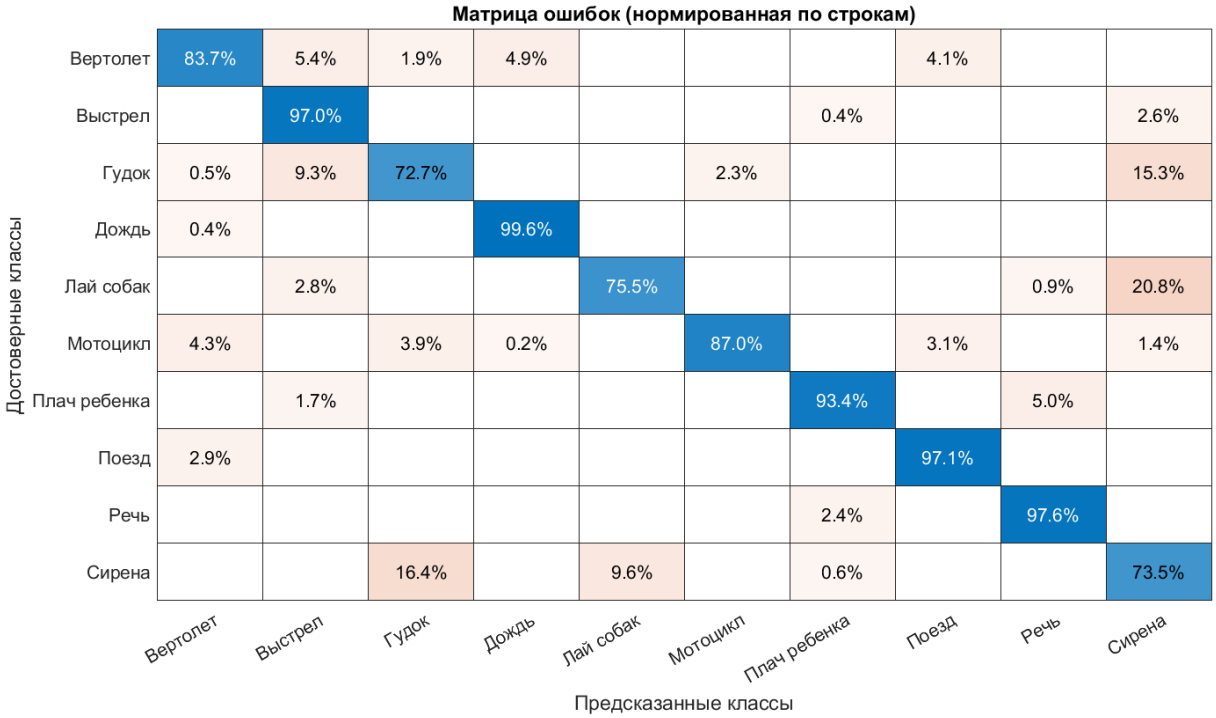


Рисунок 8 – Нормированная матрица ошибок для модели Mel/300

Общие результаты для модели Mel/300:
– Accuracy: 89,19 %;
– Macro F1: 86,88 %;
– Weighted F1: 89,21 %.

Видно, что все общие показатели немного, но увеличились по сравнению с результатами обучения модели STFT/300. Модель Mel/300 уменьшила разницу качества определения отдельного класса, что делает ее более удачной для оптимальной классификации.

Следующий эксперимент аналогичен предыдущему, но для входных данных используются Барк-спектрограммы. Такую сборку модели можно обозначить как Bark/300. На рисунке 9 показана нормированная матрица ошибок для модели Bark/300. Заметно снижение точности определения проблемных классов, хотя данные результаты все еще лучше некоторых, чем у модели STFT/300.

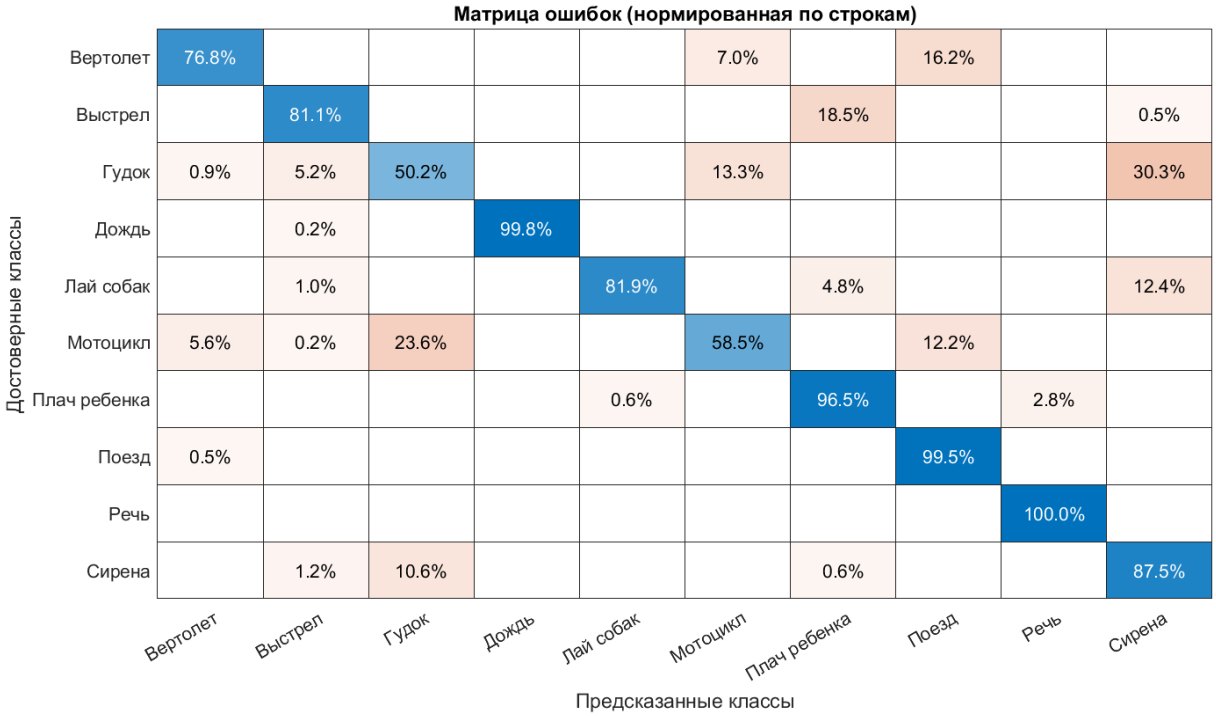


Рисунок 9 – Нормированная матрица ошибок для модели Bark/300

Общие результаты для модели Bark/300:
– Accuracy: 83,83 %;
– Macro F1: 82,29 %;
– Weighted F1: 83,82 %.

Видно, что общие показатели на 2-4% уменьшились, по сравнению с моделью STFT/300. Таким образом, модель Bark/300 в целом справляется с задачей распознавания немного хуже моделей STFT/300 и Mel/300, если рассматривать нишу с одной шириной временного интервала.

Для дополнительной проверки был проведен еще один эксперимент при количестве временных кадров 200 и мел-спектрограмме. Такую сборку модели можно обозначить как Mel/200. На рисунке 10 показана нормированная матрица ошибок для модели Mel/200. Видно, что она имеет такую же проблему с «Лаем собак», как и STFT/200, однако хоть что-то она все же определила верно. Более того модель полностью спутала данный класс с «Сиреной».

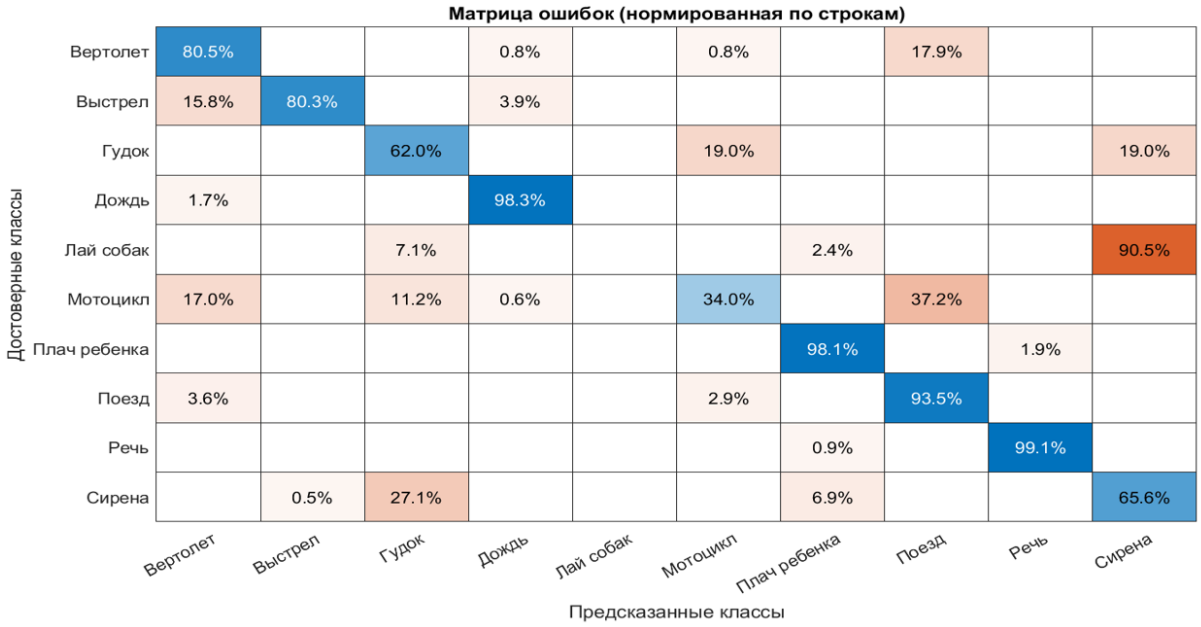


Рисунок 10 – Нормированная матрица ошибок для модели Mel/200

Общие результаты для модели Mel/200:
– Accuracy: 82,51 %;
– Macro F1: 72,61 %;
– Weighted F1: 79,69 %.

Видно, что общие показатели чуть выше, чем у STFT/200, однако из-за критической ошибки определения класса «Лай собак» такую модель тоже нельзя использовать на практике.

Для оценки эффективности применения разных параметров обработки акустических сигналов при использовании их в качестве входных данных для классификации шумов с помощью CRNN следует определить самую оптимальную модель из рассмотренных выше. Все значимые метрики показаны в сводной таблице 1, где полужирным начертанием выделены максимальные значения среди всех моделей, а курсивом – минимальные.

Результаты проведённых экспериментов позволяют утверждать, что использование Мел-спектрограммы в сочетании с временным окном 0,2 секунды (300 кадров в минуту) является наиболее оптимальным решением для обучения сверточно- рекуррентной нейронной сети в задаче классификации шумов. Данная конфигурация (Mel/300) продемонстрировала наивысшие интегральные показатели качества и стала лидером по четырём метрикам F1 из десяти. Это указывает на способность модели эффективно извлекать информативные признаки спектральной структуры шума при сохранении устойчивости к межклассовым пересечениям.

Можно проанализировать работоспособность модели на отдельных аудиодорожках и сравнить полученные результаты предсказаний с истинными, визуально оценив характер допущенных ошибок. На рисунке 11 показаны временные диаграммы успешной модели Mel/300 для 2 разных аудиофайлов, где слева представлено достоверное временное распределение по классам, а справа – предсказанное.

По диаграммам видно, что модель корректно определяет большую часть сигналов, однако помимо этого часто ложно срабатывает всплесками, которые составляют буквально 3-7 кадров, что меньше 1 секунды. Чтобы уменьшить данный «шум», можно ввести сглаживание в постобработку результатов работы обученной сети.

Таблица 1 – Метрики качества работы обученных моделей

Метрика		**Модели (по способу обработки входных данных)**					
		STFT/600	**STFT/300**	**Mel/300**	**Bark/300**	**STFT/200**	**Mel/200**
F1, %	Вертолет	68,5	74,1	**87,3**	82,8	74,4	*30,2*
	Выстрел	90,5	**94,8**	89,2	*85,7*	88,7	86,0
	Гудок	*7,9*	52,3	**69,5**	45,5	53,8	68,2
	Дождь	98,9	99,8	98,0	**99,9**	*97,7*	**99,9**
	Лай собак	80,5	77,4	73,7	**89,1**	*0*	9,8
	Мотоцикл	73,3	85,4	**92,5**	69,1	*47,0*	84,2
	Плач ребенка	98,8	**99,1**	94,6	*91,2*	94,9	95,2
	Поезд	73,7	78,3	**91,8**	77,9	*60,4*	76,7
	Речь	98,9	**99,7**	*96,6*	98,8	98,7	99,2
	Сирена	*75,6*	79,7	75,6	**82,9**	*67,1*	76,7
Accuracy, %		82,7	87,6	**89,2**	83,8	*76,4*	82,5
Macro F1, %		76,7	84,1	**86,9**	82,3	*68,3*	72,6
Weighted F1, %		81,2	87,1	**89,2**	83,8	*75,4*	79,7

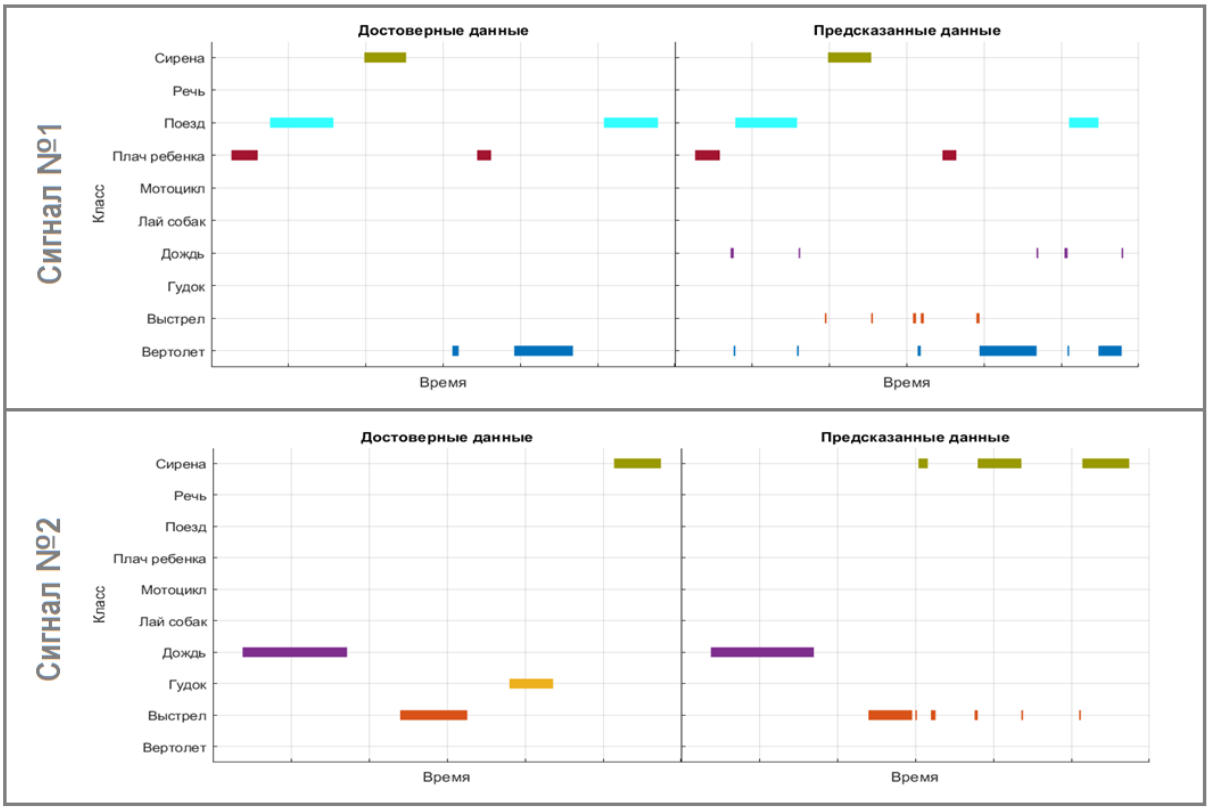


Рисунок 11 – Временная диаграмма распределения классов для модели Mel/300

Для сравнения, следующая по эффективности модель (STFT/300) показала результаты всего на 2% ниже, однако продемонстрировала меньшее количество лидерских

значений по F1 (3 из 10). Это подтверждает, что хотя использование стандартного преобразования Фурье также может быть продуктивным, Мел-спектрограмма лучше адаптирована к специфике восприятия звуков человеком и позволяет CRNN моделям эффективнее выделять значимые акустические паттерны.

Наименее успешным вариантом оказалась модель STFT/200, показавшая пять худших значений F1 из десяти и минимальные суммарные показатели. Это демонстрирует, что сокращённое временное окно (0,2 с против 0,3 с) ограничивает модель в способности улавливать длительные закономерности и формировать устойчивые временно-частотные представления.

Таким образом, комбинация Мел-спектрограммы и временного окна в 0,2 секунды обеспечивает оптимальный баланс между детализацией частотных характеристик и достаточной временной протяжённостью контекста, позволяя достичь высокой точности распознавания даже в условиях акустической вариативности.

Заключение

Результат исследования подтверждает критическое влияние на эффективность обучения нейронных сетей в задачах классификации шумов таких параметров как длина кадра классификации и формат представления акустического события. Было установлено, что оптимальной комбинацией является использование Мел-спектрограммы с длиной кадра 0,2 секунды, обеспечивающей наилучший баланс между временной и частотной разрешающей способностью. Эта конфигурация позволила построенной модели CRNN продемонстрировать максимальные суммарные показатели качества и достичь лидирующих значений по метрикам F1, что свидетельствует о высокой устойчивости к межклассовым пересечениям и способности к точному выделению акустических паттернов.

Итоги работы показывают наличие различий в методе распознавания шумов и иных звуков, например, речи, так как шумовые сигналы обладают высокой вариативностью и сложной временно-частотной структурой, что делает выбор параметров предобработки отличным от других. Правильная настройка предобработки – длины кадра и изображения спектрограммы – позволяет существенно повысить эффективность нейронных сетей без необходимости усложнения архитектуры.

Таким образом, проведённый эксперимент не только подтвердил преимущество Мел-спектрограмм в сочетании с временным окном 0,2 секунды, но и подчеркнул значимость поиска оптимальных параметров спектрального анализа как ключевого этапа подготовки аудиоданных для нейросетевых моделей.

Список использованных источников

1. Липилин Д. А. Оценка качества городской среды с применением геоинформационных систем на примере Московского микрорайона города Краснодара / Липилин Д. А., Евтушенко Д. Д. // Региональные геосистемы. - 2022. - Т. 46, N. 2. - С. 223-240.

2. Пердебаева Г. Д. Шумовое загрязнение как одна из экологических проблем современного города // Теория и практика современной науки. - 2022. - N. 8 (86). - С. 94-97.

3. Montenegro A. L. Streets classification models by urban features for road traffic noise estimation / Montenegro, A. L., Rey-Gozalo, G., Arenas, J. P., Suárez, E. // Science of The Total Environment. - 2024. - Т. 932. - С. 173005.

4. Степура Д. В. Спектральные характеристики сигналов для их классификации по уровню раздражительности методами машинного обучения / Д. В. Степура, Л. Г. Стаценко, А. Ю. Родионов // Известия СПбГЭТУ ЛЭТИ. - 2024. - Т. 17, N 4. - С. 53-60. - DOI 10.32603/2071-8985-2024-17-4-53-60. - EDN TGYXBI.

5. Nogueira A. F. R. et al. Sound classification and processing of urban environments: A systematic literature review / Nogueira, A. F. R., Oliveira, H. S., Machado, J. J., Tavares, J. M. R. // Sensors. - 2022. - Т. 22, N. 22. - С. 8608.

6. Grumiaux P. A. et al. A survey of sound source localization with deep learning methods // The Journal of the Acoustical Society of America. - 2022. - Т. 152, N. 1. - С. 107-151.

7. Hossain M. S. Environment classification for urban big data using deep learning / Hossain M. S., Muhammad G. // IEEE Communications Magazine. - 2018. - Т. 56, N.11. - С. 44-50.

8. Adavanne S. “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks” / Adavanne S., Politis A., Nikunen J., and Virtanen T. // IEEE J. Sel. Top. Signal Process. - Т. 13, N 1, - С. 34-48, 2019.

9. Mesaros A. Joint measurement of localization and detection of sound events / Mesaros A., Adavanne S., Politis A., Heittola T., Virtanen T. // 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). - IEEE, 2019. - С. 333-337.

References

1. Lipilin D. A., Yevtushenko D. D. Assessment of the quality of the urban environment using geoinformation systems on the example of the Moscow microdistrict of Krasnodar // Regional geosystems. - 2022. - Vol. 46, - N. 2. - P. 223-240.

2. Perdebaeva G. D. Noise pollution as one of the environmental problems of a modern city //Theory and practice of modern science. - 2022. - N. 8 (86). - P. 94-97.

3. Montenegro A. L. Models of street classification by urban features for assessing traffic noise / Montenegro A. L., Rey-Gozalo G., Arenas J. P., Suarez E. // Science of the environment in general. - 2024. - Vol. 932. - p. 173005.

4. Stepura D. V. Spectral Characteristics of signals for their classification by the level of irritability by machine learning methods / D. V. Stepura, L. G. Statsenko, A. Yu. Rodionov // Izvestiya SPbSETU LETI. - 2024. - Vol. 17, N. 4. - P. 53-60. - DOI 10.32603/2071-8985-2024-17-4-53-60. - ED. TGYXBI.

5. Nogueira A. F. R. et al. Sound classification and processing of urban environments: A systematic literature review / Nogueira, A. F. R., Oliveira, H. S., Machado, J. J., Tavares, J. M. R. //Sensors. - 2022. - T. 22. - N. 22. - P. 8608.

6. Grumiaux P. A. et al. A survey of sound source localization with deep learning methods //The Journal of the Acoustical Society of America. - 2022. - Vol. 152, N. 1. - P. 107-151.

7. Hossain M. S. Environment classification for urban big data using deep learning / Hossain M. S., Muhammad G. //IEEE Communications Magazine. - 2018. - Vol. 56, N.11. - P. 44-50.

8. Adavanne S. “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks” / Adavanne S., Politis A., Nikunen J., and Virtanen T. // IEEE J. Sel. Top. Signal Process. - Vol. 13, N 1. - P. 34-48, 2019.

9. Mesaros A. Joint measurement of localization and detection of sound events / Mesaros A., Adavanne S., Politis A., Heittola T., Virtanen T. //2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). - IEEE, 2019. - P. 333-337.