ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ БЮДЖЕТНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ**

**РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**(ФИНАНСОВЫЙ УНИВЕРСИТЕТ)**

Департамент анализа данных

и машинного обучения

***Дисциплина: «Технологии анализа данных и машинного обучения»***

*Направление подготовки: «Прикладная информатика»*

*Профиль: «ИТ-сервисы и технологии обработки данных в экономике и финансах»*

*Факультет информационных технологий и анализа больших данных*

*Форма обучения очная*

*Учебный 2022/2023 год, 6 семестр*

**Курсовая работа на тему:**

«Выявление речевых дефектов на основе текстов, аудированных с различением звуков»

*Выполнил:*

студент группы ПИ20-5

Макаров М. И.

*Научный руководитель:*

к.ф.-м.н., доцент Попов В.Г.

**Москва 2023**

Оглавление

[введение 3](#_Toc135214178)

[Основная часть 6](#_Toc135214179)

[Глава 1. Описание предметной области и алгоритма реализации 6](#_Toc135214180)

[1.1. Описание предметной области и ее особенности 6](#_Toc135214181)

[1.2. Метод решения анализа звукового сигнала 7](#_Toc135214182)

[Глава 2. Реализация модели. Исследование речевых дефектов. 12](#_Toc135214183)

[2.1. Набор данных 12](#_Toc135214184)

[2.2. Предобработка данных. 15](#_Toc135214185)

[2.3. Архитектура. Функция ошибки. Оптимизатор. Общая логика. 15](#_Toc135214186)

[2.4. Модель 0. Начальная 17](#_Toc135214187)

[2.5. Модель 1. Функция активации LeakyReLU. 20](#_Toc135214188)

[2.6. Модель 2. Функция активации LeakyReLU. Изменение kernel\_size и padding для 2 счерточного слоя. Увеличение Dropout. 23](#_Toc135214189)

[2.7. Модель 3. Функция активации LeakyReLU. Уменьшение слоев. Увеличение dropout 26](#_Toc135214190)

[2.8. Сравнение моделей. 28](#_Toc135214191)

[заключение 29](#_Toc135214192)

[Список литературы 30](#_Toc135214193)

[ПРИЛОЖЕНИЕ 1. ИСХОДНЫЙ КОД. 31](#_Toc135214194)

# введение

Искусственный интеллект (далее по тексту – ИИ) в настоящее время становится все более востребованным и в некоторых сферах жизни незаменим.

Это связано с тем, что ИИ способен решать задачи, для которых требуется высокая скорость обработки больших объемов информации, точность и предсказуемость. Примерами сфер, где ИИ уже применяется, являются медицина, финансы, транспорт, образование, наука, производство, розничная торговля и другие. Например, в медицине ИИ помогает диагностировать и лечить заболевания, снижать вероятность ошибок в диагностике и лечении, оптимизировать медицинские процессы и т.д.

Также следует отметить, что ИИ уже сейчас используется для решения задач, которые ранее не могли быть решены с помощью классических методов. Например, в науке ИИ применяется для анализа геномов, создания новых материалов, разработки новых лекарств и т.д.

Таким образом, ИИ уже сейчас оказывает значительное влияние на сферы жизни и, скорее всего, в будущем его применение будет только расширяться и углубляться.

Одной из социальных проблем в развитии ребёнка (так и взрослого человека. Далее по тексту будет упоминаться только ребёнок, поскольку логопедической проблематикой занимаются, как правило, в детстве) является его речь. Узкие специалисты – логопеды способны выявить и помочь исправить проблемы речи. Но они порой требуют значительные затраты времени, т.к. для выявления логопедической или дефектологической проблемы необходимо иметь специальные познания в дефектологии и проводить значительное количество экспериментальных занятий с ребёнком, но это также не исключает субъективных ошибок. В том числе, чтобы получить необходимые навыки логопеду приходится посвятить большую часть времени на приобретение опыта по выявлению определённых отклонений в речи ребёнка.

В настоящей работе поставлена цель – начать решение одной из проблем в логопедии: выявление речевых дефектов на основе текстов, аудированных с различением звуков с помощью нейросети.

Разработка нейросети, которая сможет определять ошибки звукопроизношения, позволит, во-первых, родителям самостоятельно выявить у ребёнка дефекты речи и только после этого обращаться к специалистам для их устранения или повторной диагностики, и во-вторых, специалисты-логопеды смогут использовать нейросеть как для самостоятельного обучения и получения практических навыков, так и минимизировать затраты времени для диагностики речи ребёнка, которая порой может проводиться в несколько этапов продолжительное время.

Аналогов ИИ на основе нейросети для выявления речевых дефектов нет. Существуют схожие нейросети.

Например, в сети интернет представлены нейросети, позволяющие классифицировать аудиофайлы (аудиодорожки песен) по жанрам музыки. Которые основываются на спектральных характеристиках аудиосигнала, таких как частоты и интенсивности звуковых волн, а также на временных характеристиках, таких как длительность и ритм.

Разрабатываемая нейросеть также основываются на спектральных характеристиках аудиосигнала, но в отличие от нейросети по жанрам музыки, использует Mel-frequency cepstral coefficients (далее по тексту - MFCC) - спектральные коэффициенты, которые позволяют представить спектрограмму аудиофайла в виде вектора признаков. Данная пред обратка используется, в частности, для распознавания речи, идентификации, классификации и других задач.

Несмотря на всё большее внедрение в жизнь ИИ, доступные исследования по оптимальной настойке данной нейросети в свободном доступе отсутствуют. Либо необходимо проводить десятки тысяч экспериментов, что требует большого количества вычислительных мощностей, либо вручную проводить обработку данных для поиска закономерностей.

Таким образом, для решения поставленной цели необходимо решить следующие задачи:

- определить оптимальную настройку параметров обучения;

- собрать набор данных для анализа;

- определить оптимальные параметры для получения MFCC;

- провести пред обратку данных;

- разработать программный код нейросети, которая определит наличие ошибки звукопроизношения в различных словах (для упрощения процесса обучения, будет взят только определенный перечень звуков - Ж, Р, Л, Ш).

Работа состоит из введения, теоретической и практической части, заключения, библиографии.

# Основная часть

## Описание предметной области и алгоритма реализации

### Описание предметной области и ее особенности

В логопедической практике рассматривается большой перечень звуков, в которых пациенты допускают ошибки произношения. В основном работа проводится с детьми, так как именно в дошкольном возрасте формируется правильное произношение звуков и развивается речь. Для выявления ошибок звукопроизношения используются специальные тесты и методики, которые позволяют определить недостатки в речевом аппарате и улучшить произношение звуков. Ошибки могут быть связаны как с внутренними факторами (биологические причины), так и с внешними (социальные).

Выделяются следующие виды нарушенного звукопроизношения:

* искажения, ненормированное, не встречающееся в речи произнесение звуков;
* замены, использование вместо звука, отсутствующего в речи, другого звука того же языка;
* смешения, недостаточная дифференциация двух близких по звучанию или по артикуляции звуков при их произношении;
* пропуски, отсутствие звука вследствие несформированности артикуляционной позы или вследствие нарушений фонематического слуха;

При диагностике, специалисты подбирают материал таким способом, чтобы каждый из обследуемых звуков находился в начале, в середине и в конце слова, так как в разных положениях звук произносится по-разному.

Также при подборе материала нужно понимать, что по законам русской фонетики звонкие согласные звуки в конце слов и в середине, если за ними идет другой глухой согласный, оглушается, т.е. произносятся как соответствующие им парные глухие звуки [1].

Вывод: для корректного обучения модели, необходимо использовать как можно больше различных видов нарушения и различные расположения звуков в словах.

### Метод решения анализа звукового сигнала

Речевые сигналы, являющиеся звуковыми волнами, представляют колебания давления, которые распространяются воздухом. Для анализа и обработки этих сигналов в цифровых системах необходимо преобразовать их из акустической формы в цифровую. Поэтому в данном контексте мы будем предполагать, что речевые сигналы были захвачены микрофоном и преобразованы в цифровую форму.

Далее речевой сигнал преобразуется в последовательность чисел, которая отображает относительное давление воздуха в определенный момент времени. Этот процесс называется импульсно-кодовой модуляцией (ИКМ). Как правило, ИКМ представляет собой последовательность дискретных значений, которые определяются двумя факторами: частотой дискретизации, то есть шагом по времени между значениями, и точностью, и распределением амплитуд, которые отражают относительное давление воздуха в определенный момент времени.

Формат wav-файла является наиболее распространенным для хранения звуковых сигналов. Он представляет временную последовательность с точностью до 16 или 32 бит в виде целого числа, мю-закона или с плавающей запятой. Частота дискретизации может колебаться от 8 до 384 кГц. Обычно файлы не сжимаются, поэтому для длительных записей может потребоваться много места на диске. Например, один час монофонической записи с частотой дискретизации 44,1 кГц занимает около 160 МБ дискового пространства [2].

1.2.1 Спектрограмма

Мы можем легко понимать, что означает высокий или низкий звук. Высота тона связана с частотой звука, который мы воспринимаем. Спектр Фурье сигнала показывает, какие частоты присутствуют в сигнале. Поэтому спектр - это удобная область для анализа сигналов, потому что мы можем визуально исследовать их содержание.

Однако на практике мы имеем дело со сигналами, которые записаны в дискретном времени. Следовательно, мы используем дискретное преобразование Фурье для перевода сигнала из временной области в частотную. Он отображает сигнал длины в комплексное представление частотной области из N коэффициентов как:

Если имеются только действительные входные данные, то мы можем сохранить только N уникальных частотных компонент, так как положительные и отрицательные компоненты являются комплексно-сопряженными друг другу. Однако, визуализация спектров с комплексными значениями представляет трудности. Для решения этой проблемы мы можем построить спектр магнитуды |Хk| или спектр мощности |Хk|^2. Однако, из-за больших различий в диапазонах разных частот, эти представления не могут полностью передать всю информацию.

Так как речевые сигналы не являются стационарными, то представление их спектра в частотной области даст нам среднее значение всех фонем в произносимом предложении. Однако в некоторых случаях мы хотели бы увидеть спектр каждой фонемы в отдельности. Разбивая сигнал на более короткие сегменты, мы можем сконцентрироваться на свойствах сигнала в конкретный момент времени.

Пример спектрограммы для аудиофайла представлен на рисунке 1.

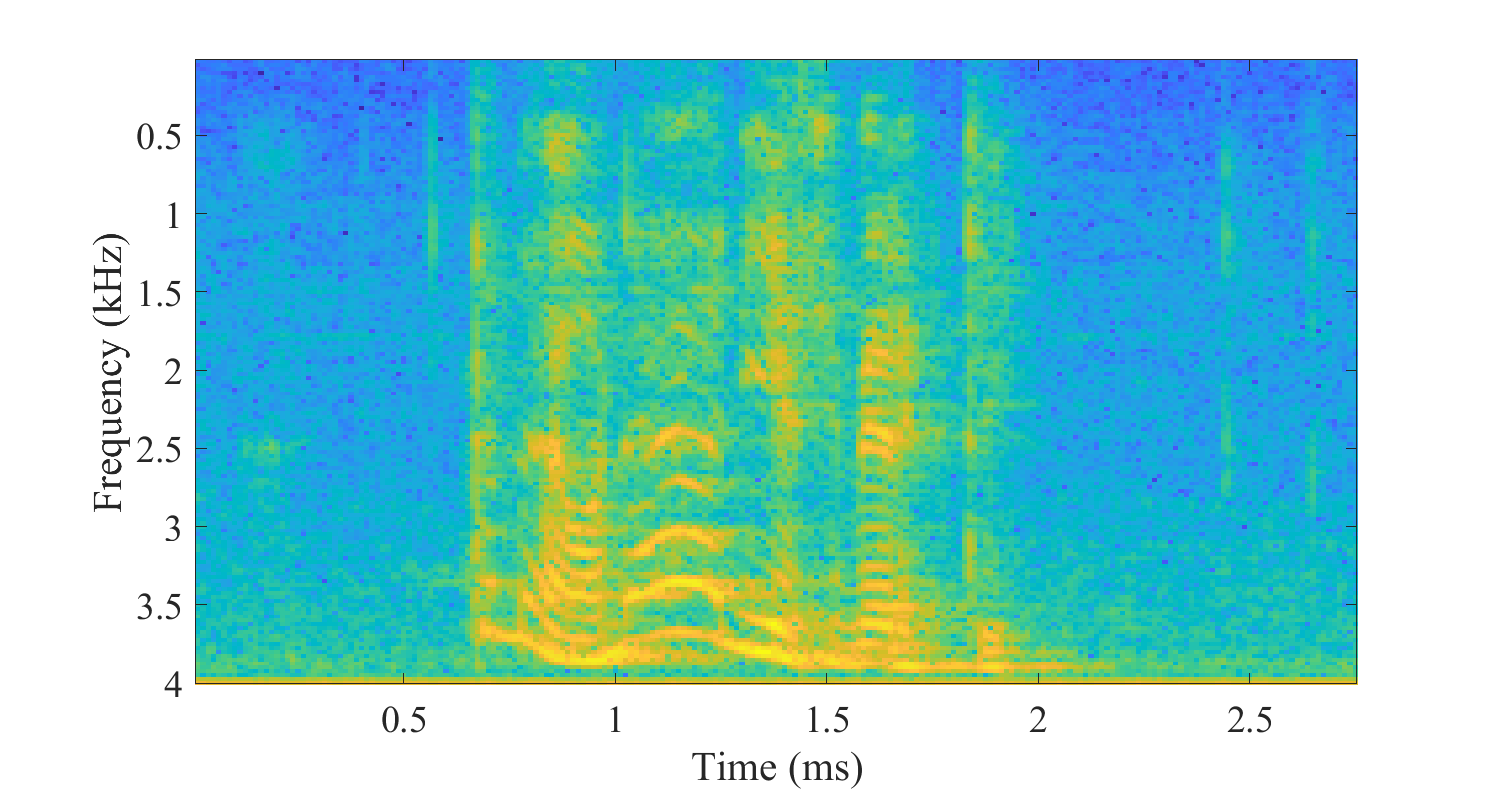


Рисунок – Спектрограмма аудиофайла

1.2.2 Кеспстр (Cepstrum)

Кепстр — это один из методов анализа периодических структур сигнала в разных масштабах, который использует преобразование Фурье. Для этого применяют дискретное преобразование Фурье (ДПФ) или дискретное косинусное преобразование (ДКП) логарифмического спектра. Кепстр получает свое название от того, что это представление является результатом последовательной обработки частотно-временных преобразований и название отражает эту игру слов. Технически, для временного сигнала x(t), кепстр определяется как:

где {⋅} представлять преобразование Фурье и {⋅} его инверсия.

Кепстр, как уже упоминалось, является комбинацией двух частотно-временных преобразований, что делает его своего рода эквивалентом временной области. На оси x кепстра находится quefrency, выраженная обычно в секундах. В кепстре более низкие частоты соответствуют медленно меняющимся характеристикам логарифмического спектра, в то время как информация о формантах находится в низкочастотной части кепстра. Тем не менее, интерпретация формантной информации в кепстре является нетривиальной задачей [3].

Пример кесптр речевого сегмента представлен на рисунке 2.

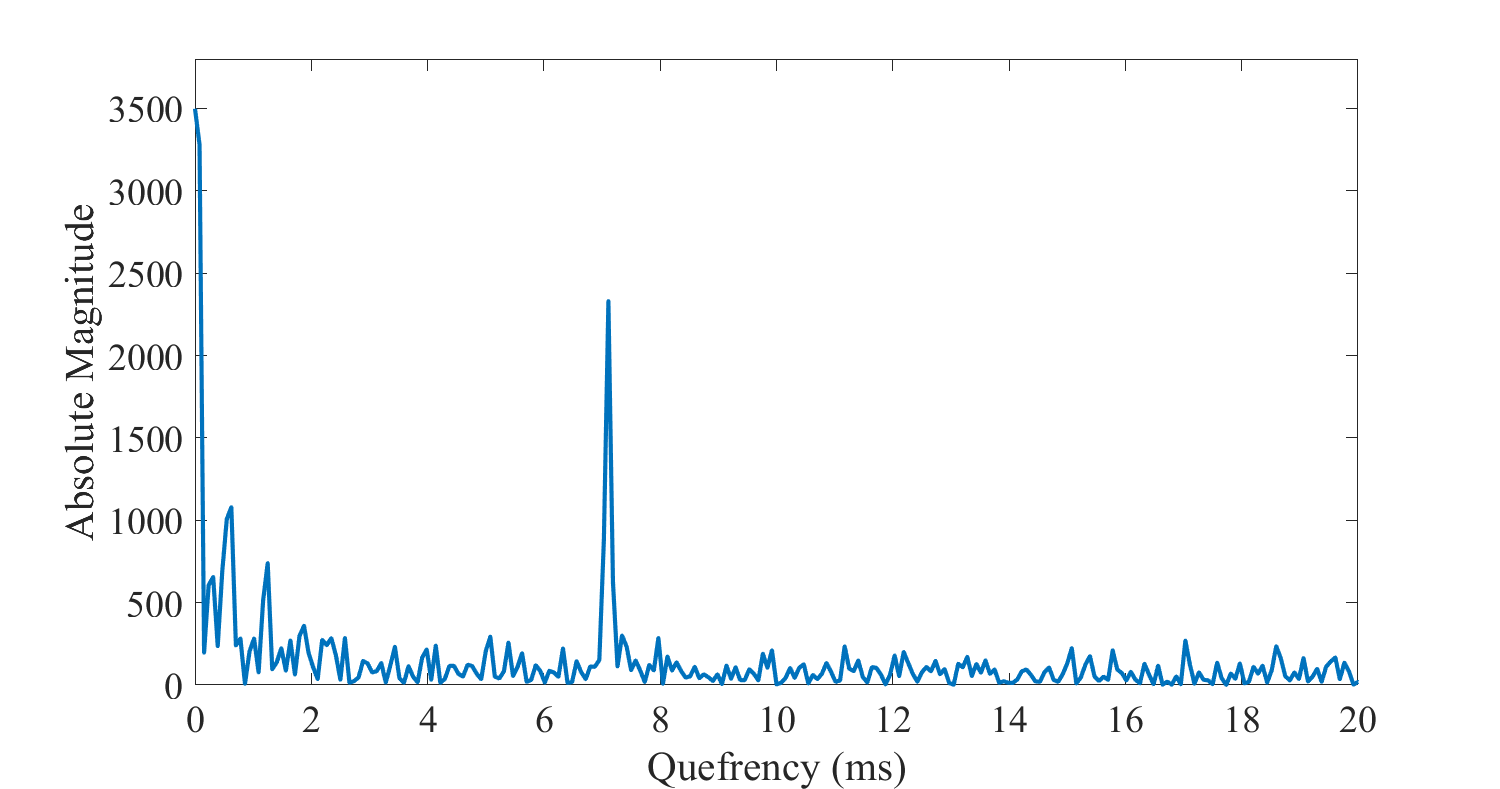


Рисунок – Кесптр аудиофайла

1.2.3 MFCC

MFCC - это сокращение от английского термина "Mel-frequency cepstral coefficients" - это набор признаков для аудио сигналов, полученных из мел-частотных кепстральных коэффициентов (Mel-frequency cepstral coefficients, MFCC).

MFCC используются в аудиообработке и распознавании речи для извлечения важных характеристик звука, которые могут быть использованы в задачах классификации или детекции звуковых событий. Они являются логарифмическими коэффициентами спектра мощности сигнала, которые были преобразованы с помощью косинусного преобразования Фурье (DFT) и мел-шкалы.

Сначала аудио сигнал делится на кадры, затем для каждого кадра вычисляется спектрограмма. Затем для каждого кадра вычисляется набор MFCC, который содержит информацию о формантах, глотательных шумах, произносительных характеристиках и других важных особенностях звукового сигнала.

MFCC имеют несколько преимуществ в сравнении с другими методами извлечения признаков, такими как спектральные признаки. Они лучше соответствуют восприятию звука человеком, так как используют мел-шкалу, а не линейную шкалу частот. Они также лучше справляются с шумом и изменениями скорости речи.

Формула вычисления MFCC может быть записана следующим образом:

где - периодограмма сигнала для j-ой частоты, M - количество коэффициентов кепстрального спектра, N - количество мел-фильтров [4].

Пример MFCC представлен на рисунке 3.

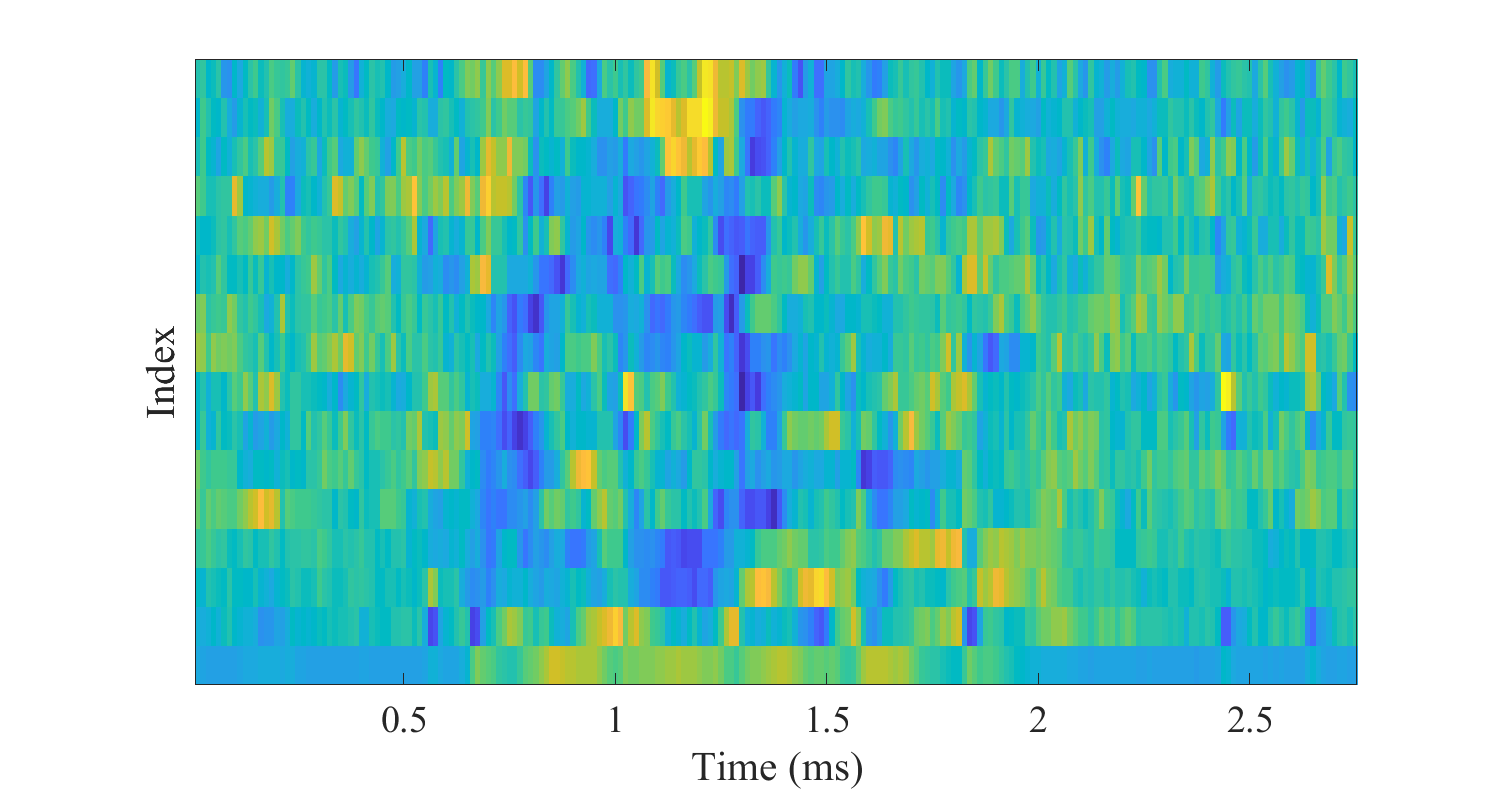


Рисунок – MFCC аудиофайла

Вывод: MFCC имеет преимущество перед кепстром и спектрограммой потому, что он учитывает особенности спектра речевых сигналов в масштабах времени и частоты. Таким образом, MFCC позволяет получить компактное представление звукового сигнала с учетом его спектральных свойств, а также он позволяет избежать проблемы с шумом, которые могут возникнуть при использовании других методов. MFCC включает в себя многократное применение оконной функции, которая позволяет снизить влияние шума на результат анализа, что может быть полезным для задач распознавания речи.

## Реализация модели. Исследование речевых дефектов.

### Набор данных

2.1.1 Способ получения данных

Исходя из выводов, сделанных в пункте 1.1, нам необходим набор аудиофайлов, каждый из который включает в себя произношение 1 слова. Файлы должны разделятся как по звукам, так и по наличию ошибки в звукопроизношении. В сети интернет подобных наборов данных найдено не было, а сбор такой информации требует большого количества времени и ресурсов, что выходит за рамки сроков курсовой работы.

В связи с этим было принято решение реализовать генерацию аудиофайлов, с использованием нейросети «Silero», которая предоставляет возможность создавать звуковой поток, на основе текста.

Нейросеть достаточно точно воспроизвела переданный ей текст. Благодаря Speech Synthesis Markup Language (далее - SSML) получилось реализовать правильные ударения в словах, которые подразумевали ошибку в звукопроизношении.

2.1.2 Алгоритм генерации данных

В качестве рассматриваемых звуков, были взяты Ж, Л, Р, Ш. Между ними есть непрямая взаимосвязь, например наиболее распространенная ошибка звукопроизношения для звука Р – замена на Л. Также имеется связь между Ж и Ш, Ш и Л. Этот набор звуков наиболее показателен с точки зрения точности, определить принадлежность слова к определенному звуку, когда происходит замена на другой рассматриваемый, вызывает трудности не только у нейросети, но и у специалистов в области дефектологии.

Структура, по которой создавались текста для нейросети, представлена в следующем виде:

1. Ж:
   1. Жаба;
   2. Ножи.
2. Л:
   1. Лук;
   2. Молоко;
   3. Стул.
3. Р:
   1. Рука;
   2. Барабан;
   3. Топор.
4. Ш:
   1. Шина;
   2. Мешок;
   3. Душ.

Структура содержит 4 звука, в каждый из которых включено по 3 слова (за исключением Ж), которые соответствуют одной из 3 возможных позиций, которые необходимо учитывать при анализе ошибки звукопроизношения. Ошибки в звукопроизношении были представлены для каждого слова в 3-4 вариациях и выбирались случайным образом, ошибки подбирались в соответствии с видами нарушений, описанных в пункте 1.1. Случайный выбор ошибки был принят с целью научить нейросеть в первую очередь определять нормальное произношение для слов, так как в реальной жизни это работает именно так. Мы знаем, как звучит одно слово и произношение в другой интерпретации сигнализирует об ошибки в звукопроизношении. Соотношение нормального звука и ошибочного 50/50.

Пример неправильного произношения для слова «лук» - вук, уук, люк, йук.

2.1.3 Представление данных

Так как все данные полностью генерируемые, было создано 3 отдельные набора с 4400 аудиофайлов для обучения (рисунок 4) и 2200 для тренировки и проверки (рисунок 5 и 6 соответственно).

Все файлы имеют разную продолжительность (от 2 до 5 секунд) и даже в файлах с произношением одного и того же слова, продолжительность различается, за счет различных голосов и пауз в начале и конце, которые также были выбраны случайным образом и реализованы с использованием SSML.



Рисунок – Обучающая выборка



Рисунок – Тренировочная выборка



Рисунок – Проверочная выборка

2.1.4 Недостаток набора данных

Так как данные были сгенерированы нейросетью, они исключают фоновые шумы, которые всегда присутствуют при записи аудиофайлов (исключения студийная запись), также в большей части набора присутствуют артефакты, они наблюдаются в начале и конце аудиозаписей (рисунок 7).

После генерации, некоторое количество аудиофайлов было прослушано специалистом логопедом, который подтвердил валидность записей и подчеркнул высокую точность в произношении ошибок.

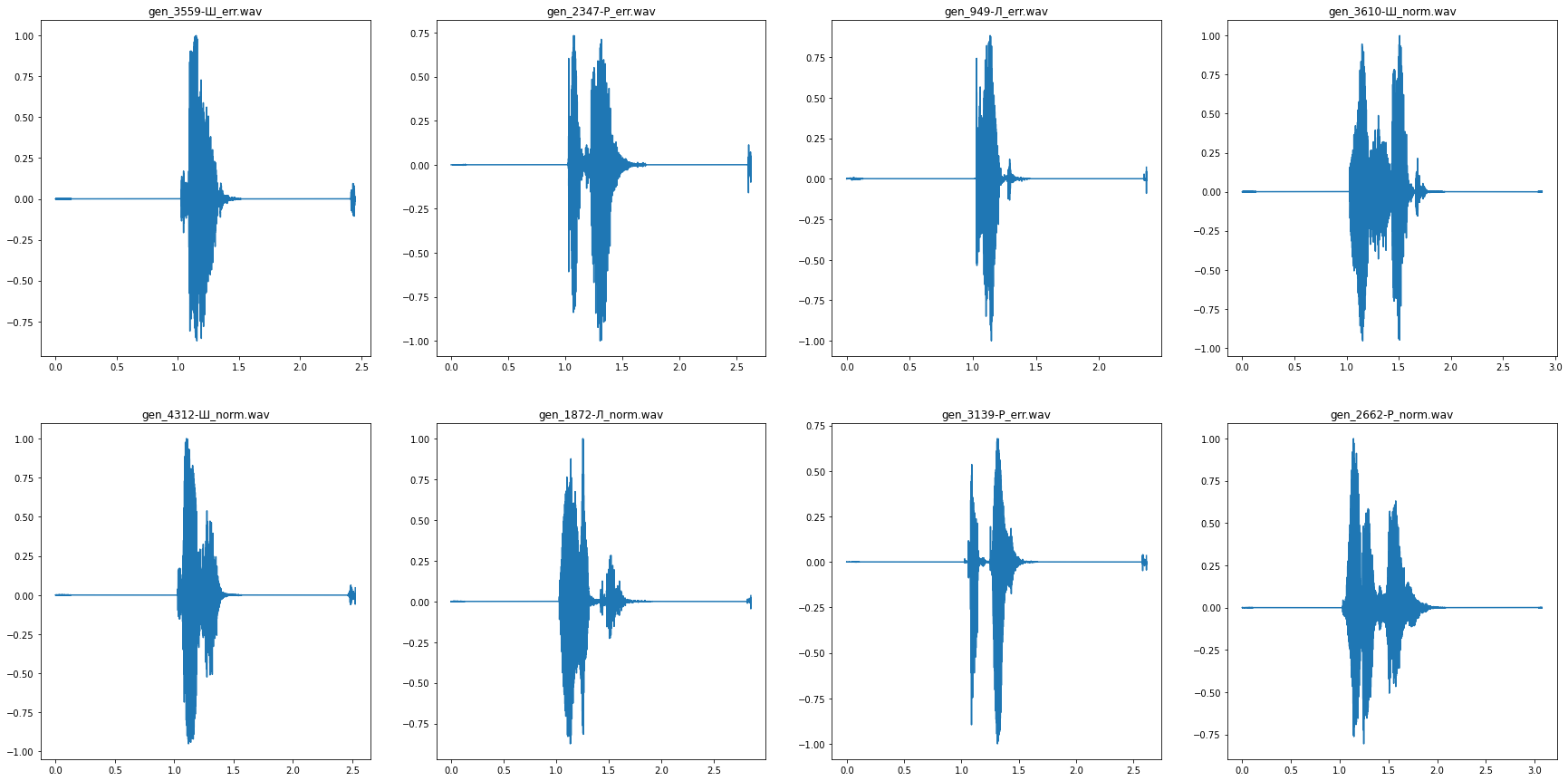


Рисунок 7 – Представление звуковой волны 3 аудиофайлов

Данные недостатки могут повлиять на итоговые результаты обучения, также стоит учитывать взаимосвязь звуков, описанную ранее, в подпункте 2.1.2.

### Предобработка данных.

В связи с различием по длительности аудиодорожек, необходимо привести данные к общей длине, после чего получить MFCC.

Параметры MFCC были выбраны следующие:

* длина кадра 6.25;
* шаг между кадрами 3.125.

По умолчанию, эти параметры равны 25 и 10 соответственно, но нам нужна максимальная точность, так как мы рассматриваем отдельные звуки в словах.

### Архитектура. Функция ошибки. Оптимизатор. Общая логика.

В процессе анализа схожих нейросетей, по анализу звукового потока с использованием MFCC, было принято решение использовать сверточную нейросеть (далее - CNN).

Архитектура модели, следующая:

1. Входной сигнал имеет форму 13 x 1465 (13 MFCC-признаков и 1465 временных отсчетов), он преобразуется в форму (-1, 1, 13, 1465) для подачи на вход сверточным слоям.
2. Сверточный слой с 32 ядрами размера (5, 11) с шагом 1 и со смещением 2 по вертикали и 5 по горизонтали, что позволяет извлечь различные признаки из спектрограммы звукового сигнала.
3. Результат слоя проходит через функцию активации ReLU для нелинейного преобразования полученных признаков.
4. Слой нормализации по батчу для борьбы с проблемой внутренней ковариации, что увеличивает стабильность и скорость обучения сети.
5. Пулинг-слой с размером окна (2, 2), уменьшающим размерность, что позволяет снизить переобучение.
6. Затем происходит свертка с 64 ядрами размера 3 с шагом 1 и со смещением 1, с последующей активацией ReLU, нормализацией по батчу и пулинг-слоем с размером окна (2, 2).
7. Полученный результат преобразуется в удобную для подачи форму на полносвязные слои.
8. Происходит применение dropout-регуляризации, который позволяет уменьшить переобучение, после чего выход проходит через полносвязный слой с 128 нейронами и функцией активации ReLU.
9. Применяется dropout-регуляризация и происходит применение полносвязного слоя с 8 выходами.
10. Использование функции активации Softmax на последнем слое позволяет получить вероятностное распределение по классам, что позволяет определить наиболее вероятный класс для данного звукового сигнала.

Таким образом, данная архитектура нейросети обладает достаточным количеством сверточных и полносвязных слоев, а также слоями нормализации и регуляризации, что делает ее эффективной для классификации звукопроизношения на основе MFCC.

Для оптимизации будет использоваться Adam, так как он показал лучшую точность на этих данных по сравнению с SGD.

В качестве функции ошибки для классификации используется категориальная кросс-энтропия.

Проверка результатов будет производиться посредством вычисления метрик эффективности, таких как:

* accuracy, доля правильных ответов;
* precision, доля ответов, которые модель определила как класс X и они действительно являются классом X;
* recall, доля ответов, которые являются классом X на самом деле и модель их определила как класс X;
* f1 score, сводная метрика, вычисляемая на основе precision и recall.

### Модель 0. Начальная

Модель обучалась 10 эпох. Изначально рост точности модели был стремительный, но затем ошибка начала уменьшаться все медленнее, и точность росла все меньше, за исключением 4 эпохи, где мы наблюдаем резкий скачек. Общее время обучения модели составило около 24 минут и точность валидационой выборки составила 89%. График изменения метрик эффективности по эпохам представлен на рисунке 7.

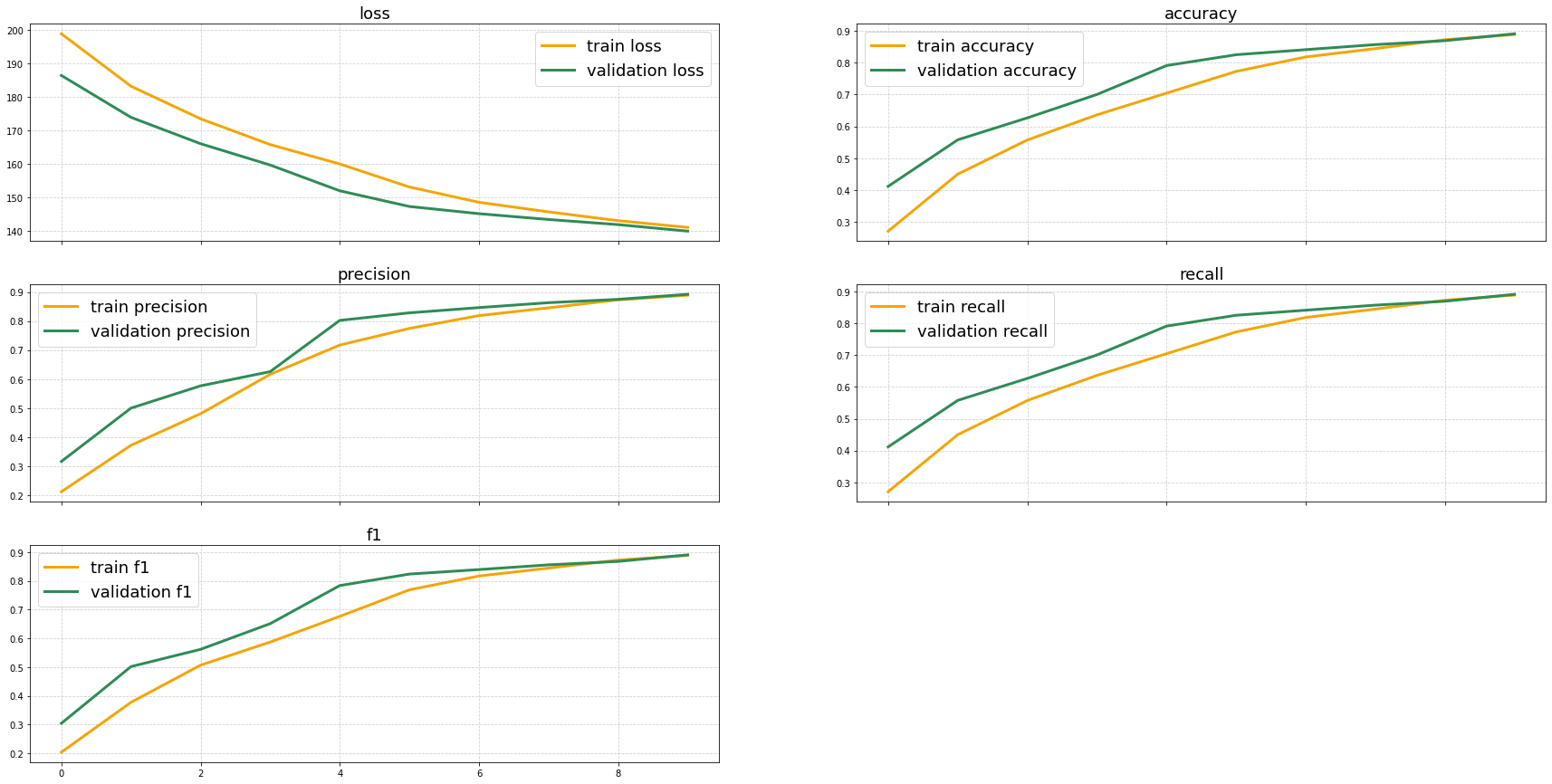


Рисунок – Изменение метрик эффективности по эпохам первоначальной модели

На графиках наглядно видно, что модель не переобучается, но близится к своему пределу, это можно заметить по графику ошибки, где валидационная выборка в начале имеет разрыв с обучающей, но в конце они сравниваются. Для более детального анализа необходимо протестировать модель на тестовом множестве и построить тепловую карту. Анализ результатов произведен с помощью функции classification\_report модуля sklearn, рисунок 8. Тепловая карта представлена на рисунке 9

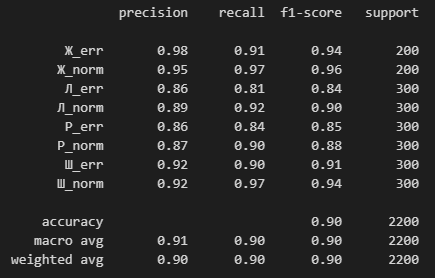


Рисунок – Отчет по модели 0

По показателям наблюдаем низкий recall по звуку Л с ошибкой.

На рисунке видно, что по звуку Л с ошибкой самый низкий recall, то есть модель ошибочно маркирует объекты этого класса, как объекты другого класса. Но в среднем метрики равны 90%. Лучше всего распознаются звуки Ж и Ш, у них метрика f1 выше 90%.

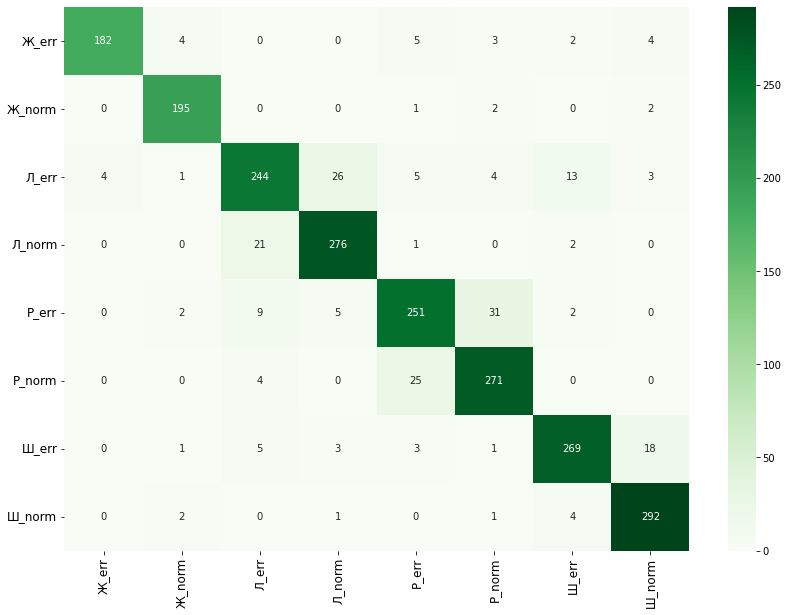


Рисунок – Тепловая карта модели 0

На тепловой карте мы видим, модель достаточно точно определяет принадлежность слов к звуку.

Около 10% на звуках без ошибки Л и Р определяется как с ошибкой и по 10% наоборот. Это связано с одним из видов нарушения искажение, когда человек пытается произнести звонкий звук, но у него не получается из-за физиологических проблем (плохой тонус мышц, дефект речевого аппарата и т.п.). Не всегда получается правильно определить нарушение в этих звуках, т.к. искажение подразумевает подражание, то есть похожий звук, а на общем уровне мощности это сложно отличить. Данная проблема может быть решена путем увеличения фреймов, получаемых с каждого кадра (параметры MFCC).

Также класс Р\_err определяется как Л\_norm и Л\_err. Тут влияет другой вид нарушения замена, т.к. Р в основном заменяется на Л и сеть делает на этом ошибки.

Связь между Л\_err и Ш\_err в теории присутствует, но скорее этот показатель нужно рассматривать как недообучение модели.

### Модель 1. Функция активации LeakyReLU. Изменение kernel\_size и padding для 1-го сверточного слоя

Одним из недостатков ReLU является возможность появления "мертвых" нейронов, когда активация ReLU становится отрицательной и не может восстановиться в положительном диапазоне. Это приводит к тому, что градиент не может протекать через этот нейрон, что может замедлить обучение и уменьшить точность сети.

LeakyReLU решает эту проблему, добавляя небольшой наклон в отрицательной области. Таким образом, при использовании LeakyReLU вместо ReLU, нейроны не могут "умереть", что может ускорить обучение и увеличить точность сети [5].

Изменение параметров kernel\_size и padding также может повлиять на характеристики сверточных фильтров и на степень сжатия данных во время свертки.

Более большие значения kernel\_size и padding могут привести к большему количеству параметров и более медленному обучению модели. Однако, увеличение kernel\_size и padding может улучшить качество выходного изображения, так как это позволяет модели выявлять более сложные и абстрактные признаки [6].

График изменения метрик на обучающей и валидационной выборке представлен на рисунке 10.

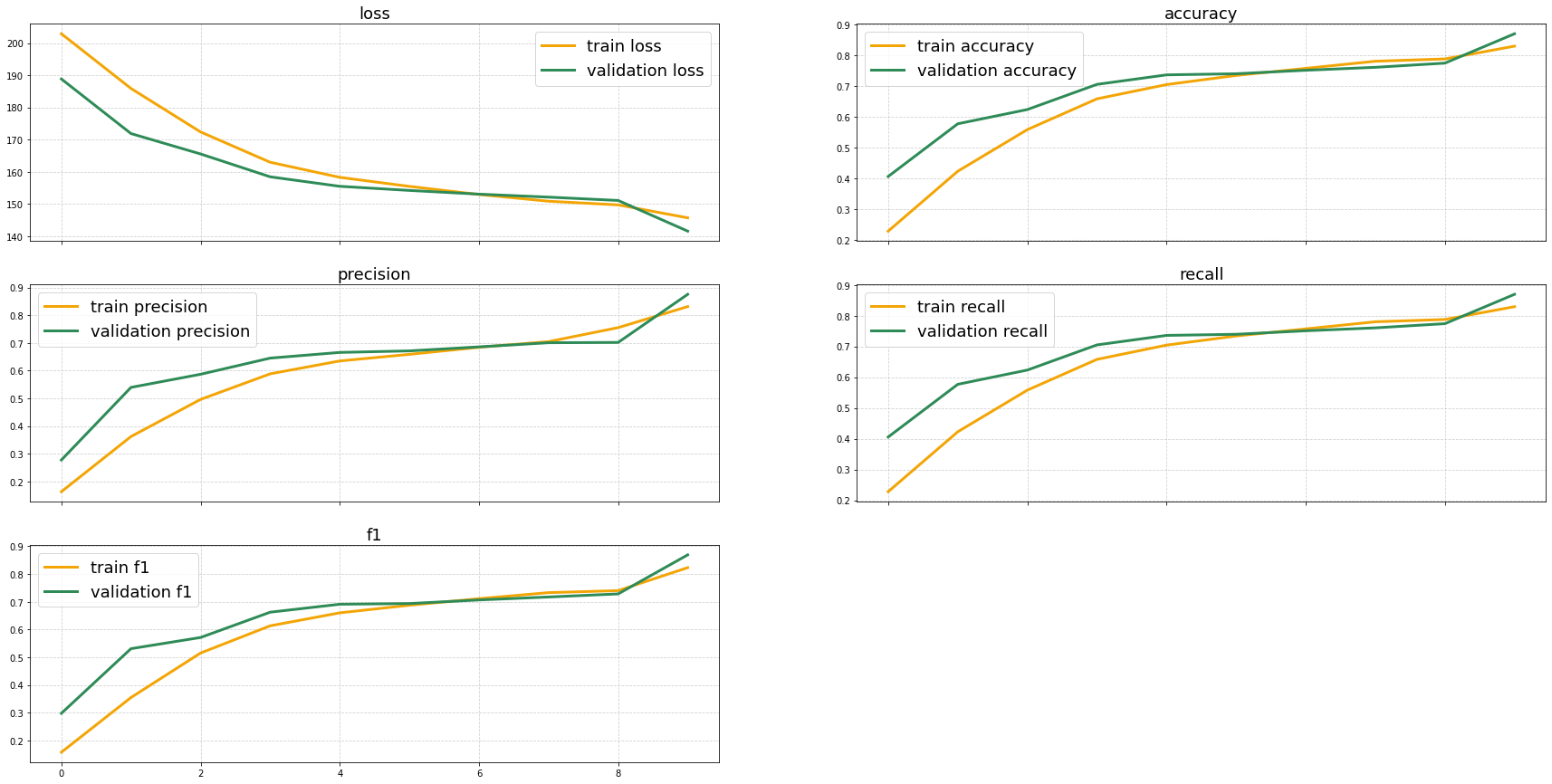


Рисунок – Изменение метрик эффективности по эпохам для модели 1

Начальной модели (далее – модель-0) потребовалось ~25 минут на обучение, и она показала точность в 89% за 10 эпох, при отсутствии явного переобучения.

Текущая модель (далее модель-1) обучилась за ~28 минут с точностью 87% за 10 эпох, но в какой-то момент она больше подстроилась под обучающую выборку.

Модель-1 показывает результаты хуже, не смотря на принятые решения по улучшению точности. На основании представленных данных сложно однозначно определить, переобучилась модель или нет, так как ошибка на тренировочной выборке продолжает снижаться, а на валидационной начинает снижаться только с 8-й эпохи. Однако, увеличение точности на тренировочной выборке и разрыв между тренировочной и валидационной точностями может свидетельствовать о том, что модель начинает переобучаться. Но на 10 эпохе модель начинает лучше обобщать данные.

Возможно, на 10 эпохе произошло снижение переобучения благодаря dropout. Также, может быть, на 10 эпохе модель была обучена на большем количестве данных или ей были предоставлены лучшие данные, что способствовало улучшению ее обобщающей способности. Еще одна возможность заключается в том, что изменение гиперпараметров модели (например, уменьшение размера сети или скорости обучения) помогло снизить переобучение и улучшить обобщающую способность модели. Отчет представлен на рисунке 11.

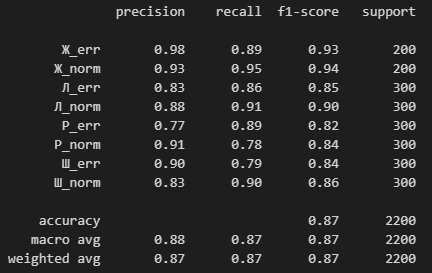


Рисунок – Отчет по модели 1

Практически по всем показателям, модель показа результаты метрик хуже, чем изначальная модель. Большое снижение (~10%) наблюдается по pricision: Ш\_norm, Р\_err; и recall: Ш\_err, Р\_norm. Далее рассмотрена тепловая карта (рисунок 12).

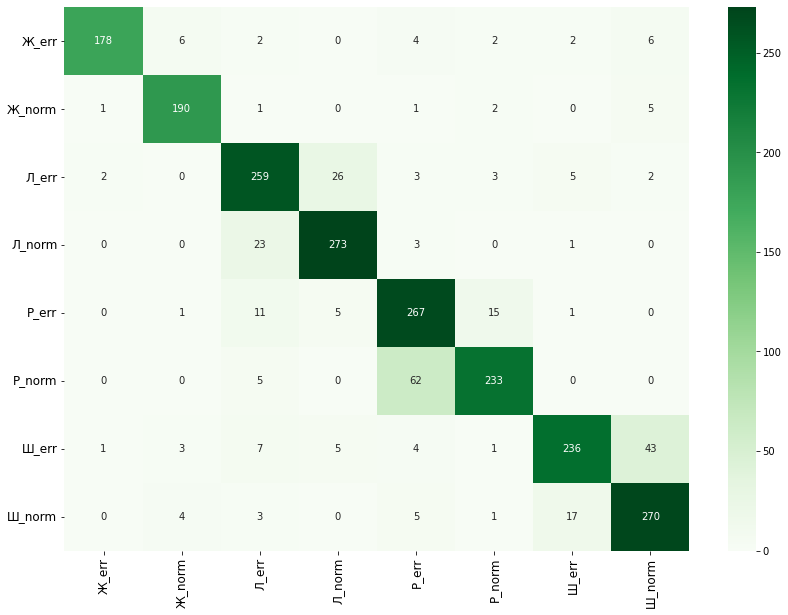


Рисунок – Тепловая карта для модели 1.

На тепловой карте мы можем заметить, что в матрицах каждого звука увеличилось количество ответов по побочной диагонали и разброс по другим звукам в сравнении с изначальной моделью.

Из положительных изменений это Л\_err, модель определила этот класс более точно и в меньшей степени определяла как Ш\_err.

Аномально высоким значением можно отметить Р\_norm - Р\_err. Как я писал ранее, ошибка в произношении звука Р чаще всего является замена.

Нельзя сказать, что данная модель стала лучше изначальной, скорее она недообучилась.

### Модель 2. Функция активации LeakyReLU. Изменение kernel\_size и padding для сверточных слоев. Увеличение Dropout.

В данном случае, изменения kernel\_size и padding в первом сверточном слое были взяты из модели-1, а во втором сверочном слое, kernel\_size (3) => (3, 7), padding (1) => (1,3).

Увеличение значения dropout помогает снизить переобучение и повысить обобщающую способность модели. Однако, если значение параметра dropout слишком большое, то сеть может начать забывать важную информацию, поскольку многие нейроны будут исключены из обучения. Это может привести к снижению точности модели на тестовых данных. Увеличение было сделано на 0,15 и составило 0,65.

В данном случае, скорость обучения была повышена в 10 раз, так как модель обучалась достаточно медленно. Число эпох было увеличено до 15, для более наглядного результата, что позволило превзойти модель\_0 по точности.

За счет повышенного dropout модель стала обучаться медленнее, но это позволило увеличить точность на 1% (не слишком большой прирост, относительно времени, которое увеличилось на ~20 минут, и того факта, что пришлось увеличить скорость обучения). График изменения метрик представлен на рисунке 13

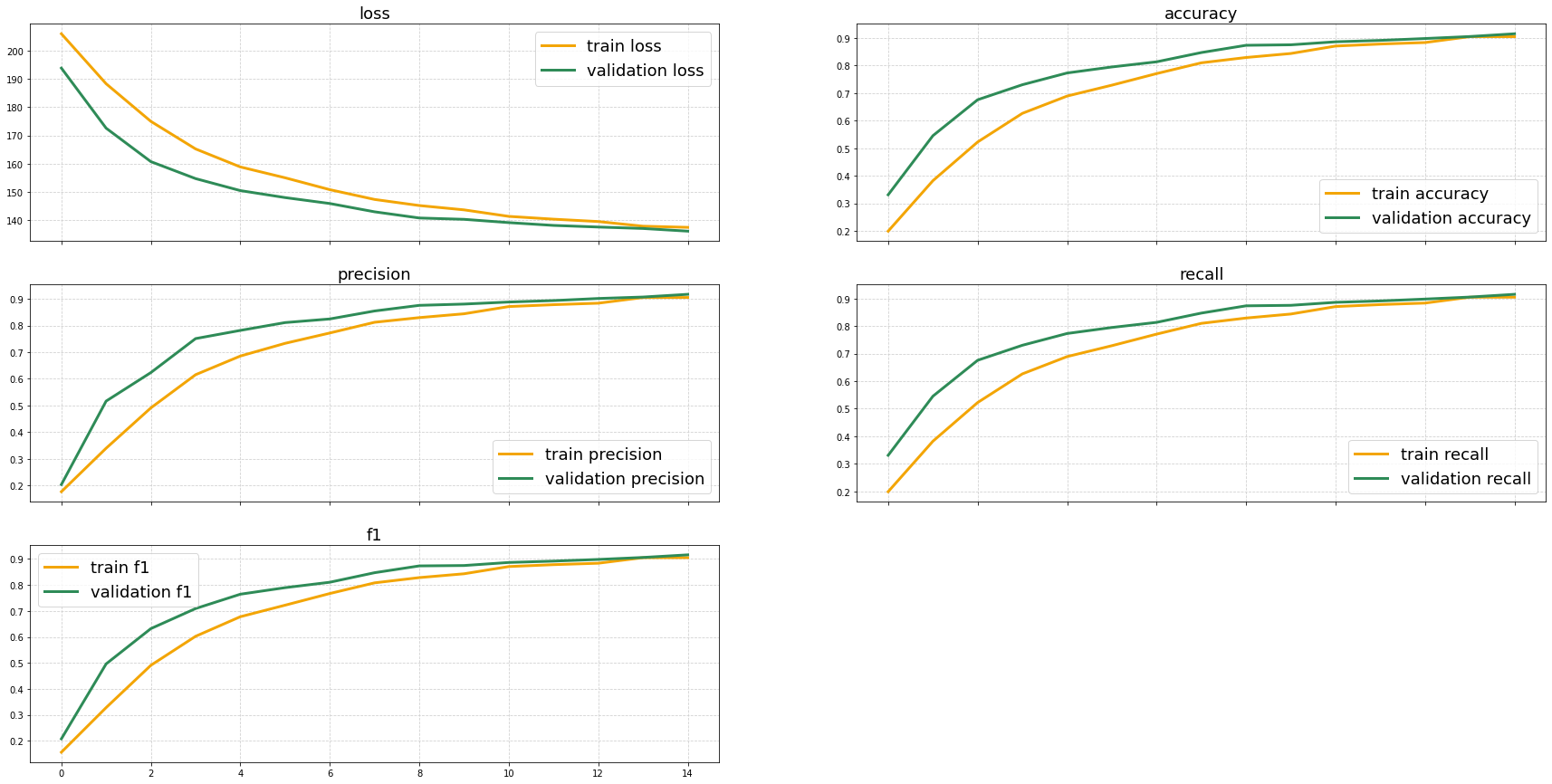


Рисунок - Изменение метрик эффективности по эпохам для модели 2

По графику precision мы видим стремительное изменение валидационной выборки на первых 4 эпохах. Падение/рост показателей по всем графикам плавный, отклоений не наблюдается. Несмотря на то, что ошибка продолжает падать, модель приближается к своему пределу. Дальнейшее обучение не имеет смысла, точность увеличится до -+1% и модель переобучится. Необходимо проанализировать отчет, он представлен на рисунке 14 и тепловую карту на рисунке 15.

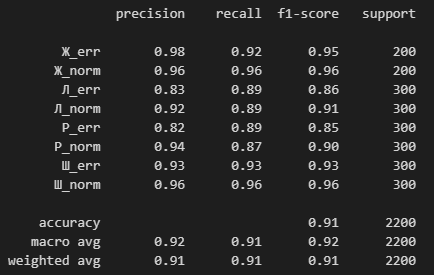


Рисунок - Отчет по модели 2

Сравнивая с моделью-0, мы видим, что по всем классам в метрике f1-score показатель увеличился или остался прежним и стали более стабильными. Но при этом precison и recall отличаю на несколько процентов.

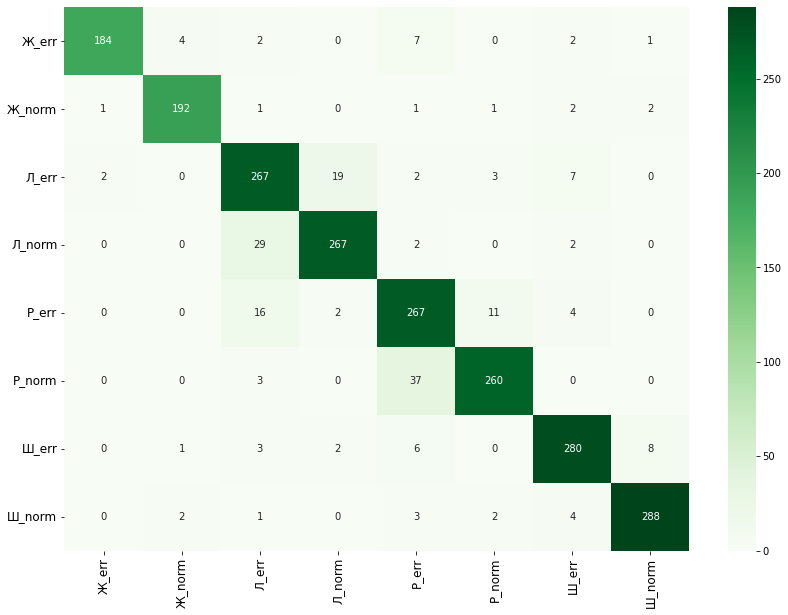


Рисунок – Тепловая карта для модели 2

На тепловой карте мы также видим минимальный разброс по звукам до 10 (исключением является Р\_err - Л\_err, в этой ситуации я считаю это нормой).

Разброс у звуков Л и Р по побочной диагонали также остался, и большая часть сфокусирована на определении нормального звука как ошибочного, в то время как на модели\_0 все было наоборот (что скорее является правильным исходом).

Модель стала лучше определять ошибочные звуки, но хуже нормальные. Можно сделать вывод, что данная модель хуже справится с реальными данными.

### Модель 3. Функция активации LeakyReLU. Уменьшение слоев. Увеличение dropout

Данная модель реализована в сторону упрощения. Убрана последовательность слоев, которая включала в себя сверточный слой, функцию активации, нормализацию батча и пулинг-слой.

Такое изменение может привести к уменьшению количества параметров в модели, что может помочь в борьбе с переобучением при обучении на маленьких датасетах. Однако, это также может уменьшить способность модели изучать более сложные зависимости в данных, что может привести к ухудшению результатов модели.

Кроме того, удаление слоя активации LeakyReLU, может уменьшить способность модели к обнаружению нелинейных зависимостей в данных.

Увеличение dropout было сделано на 0,25 и составило 0,75.

Для получения хоть каких-то сравнимых результатов модель обучалась 18 эпох, при увеличенной скорости обучения в 10 раз от изначальной, при этом время обучения уже было практически в 2 раза больше, чем для модели-0 и составило ~53 минуты. Результаты представлены на рисунке 16.

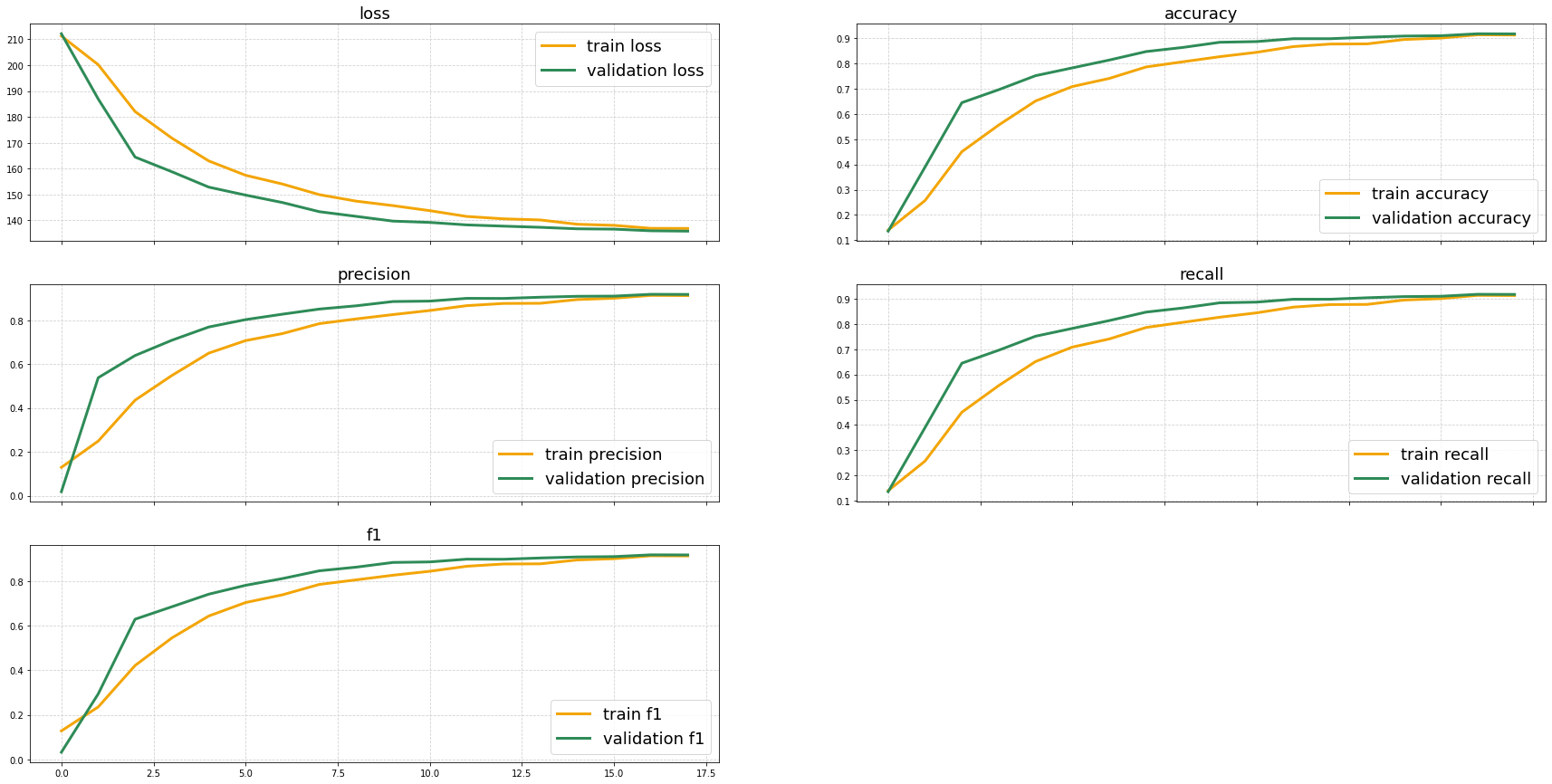


Рисунок - Изменение метрик эффективности по эпохам для модели 3

По результату мы видим не значительный прирост точность относительно изначальной модели, при более длительном обучении.

На графиках мы видим, что первая эпоха прошла не лучшим образом, ошибка валидационной выборки больше, чем тренировочной, точность ниже.

Модель тоже дошла до своего предела обучения, это видно по разнице между ошибок обучающей и валидационной выборки. Отчет по тестовой выборке представлен на рисунке 17.

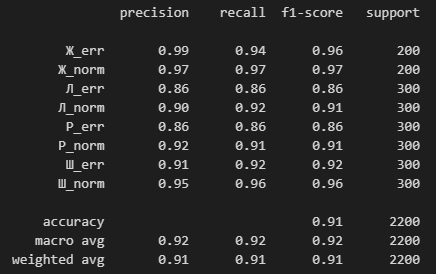


Рисунок – Отчет по модели 3

Явные улучшения наблюдаются в Р\_norm (по всем метрикам), recall Л\_err, Ж\_err. Далее представлена тепловая карта на рисунке 18.

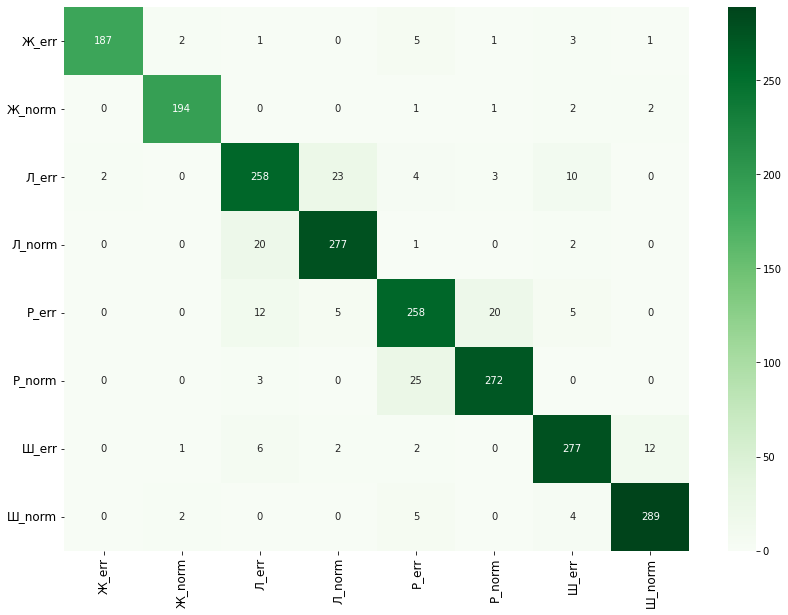


Рисунок – Тепловая карта для модели 3

Сравнивая с моделью-0, можно сказать, что данная модель по всем звукам показывает более высокую точность (если брать матрицу нормального произношения и ошибочного) и более высокую точность по классам.

### Сравнение моделей.

Сравнивая модель\_0 с другими ее архитектурами, можем сказать, что:

* модель-1 является наихудшим выбором, результаты точности ниже, чем у модели-0, а время обучения гораздо больше;
* модель-2 показа точность выше, не смотря на длительное время обучения, но она потеряла точность в определении нормальных звуков, что на практике не будет эффективным;
* модель-3 способна конкурировать с модель-0, ее точность по звуку и конкретным классам выше и лучше, окончательный выбор можно сделать только используя реальные данные, полученные путем сбора. Модель\_0, выглядит перспективнее, за счет большего количества слоев и возможности их настройки.

# заключение

В заключение к проделанной работе можно выделить следующие факт:

* LeakyReLU не показала прироста в точности в сравнении с relu для данной задачи;
* Благодаря MFCC, нейросети смогли вычленять из большого количества звуков необходимые в исследовании.
* Увеличение dropout давал прирост в точности, за счет более длительного процесса обучения;
* Каждая из представленных моделей отлично справлялась с классификацией слов по звукам, это говорит о перспективности дальнейшего исследования.

Несмотря на то, что ни одна из моделей не смогла достичь результата в 95% точности и выше, исследование считаю успешным. Так как модели обучались на условно-фиктивных данных, сгенерированных нейросетью. В этих данных могут присутствовать не точности произношения, не четкого ударения и различного рода артефакты.

Результаты данной работы могут быть применены в профессиональной деятельности специалистов логопедии и дефектологии. Это может как самим специалистам сократить время диагностирования детей, так и родителям избежать последующего наслоения нарушений речи ребенка при не своевременном обращении к специалисту.

# Список литературы

1. Т.Д. Филичева. Логопедия. Теория и практика – 2020 (609 стр.)
2. Waveform. URL: https://wiki.aalto.fi/display/ITSP/Waveform (дата обращения 02.03.2023)
3. Spectrogram and the STFT. URL: https://wiki.aalto.fi/display/ITSP/Spectrogram+and+the+STFT (дата обращения 02.03.2023)
4. Cepstrum and MFCC. URL: https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC (дата обновления 02.03. 2023)
5. Activation Functions in PyTorch. URL: https://machinelearningmastery.com/activation-functions-in-pytorch/ (дата обращения 16.04.2023)
6. Padding and Stride. URL: https://d2l.ai/chapter\_convolutional-neural-networks/padding-and-strides.html (дата обращения 16.04.2023)

# ПРИЛОЖЕНИЕ 1. ИСХОДНЫЙ КОД.

1. https://github.com/Ara-Bog/Kursovaya2/blob/master/generator/gen.ipynb
2. https://github.com/Ara-Bog/Kursovaya2/blob/master/main.ipynb