

Министерство цифрового развития, связи и массовых коммуникаций
Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

Кафедра _____ ТСиВС
Допустить к защите

Зав.каф. _____ Дроздова В.Г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Решение задачи распознавания эмоций в
человеческой речи на основе методов глубокого
обучения

Пояснительная записка

Студент _____ Ющенко А.В. /...../

Факультет _____ ИВТ _____ Группа _____ ИА-032

Руководитель _____ Лошкарёв А.В. //

Новосибирск 2024 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	3
1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ.....	7
1.1 Психологические и лингвистические аспекты выражения эмоций в речи.....	7
1.2 Обзор методов распознавания эмоций	8
1.3 Наборы данных для распознавания эмоций.....	10
1.4 Подготовка данных для моделей глубокого обучения	11
1.5 Метрики оценки качества моделей	17
2 МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ	20
2.1 Выбор данных для обучения и тестирования модели.....	20
2.2 Подготовка данных	22
2.3 Предлагаемая архитектура нейросетевой модели	24
2.4 Оценка качества модели.....	26
2.5 Описание вычислительной платформы и инструментария	27
2.6 Описание программного инструментария.....	28
3 ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ	33
3.1 Обучение и оценка эффективности модели	33
3.2 Тестирование модели на различных наборах данных.....	39
4 АНАЛИЗ РЕЗУЛЬТАТОВ РАБОТЫ	48
4.1 Оценка качества распознавания эмоций на различных наборах данных ...	48
4.2 Анализ факторов, влияющих на распознавание эмоций	48
ЗАКЛЮЧЕНИЕ	50
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	51

ВВЕДЕНИЕ

В современном мире искусственные нейронные сети (ИНС) стали неотъемлемой частью нашей повседневной жизни. Интеллектуальные системы и алгоритмы, основанные на ИНС, применяются для решения широкого спектра задач в самых разных областях — от обработки естественного языка и компьютерного зрения до управления автономными системами и поддержки принятия решений. По некоторым показателям, таким как скорость работы, точность и способность обрабатывать огромные массивы данных, ИНС во многом превосходят возможности человека. Это позволяет значительно упростить и ускорить множество процессов как в бизнесе, так и в повседневной жизни.

Одним из перспективных и активно развивающихся направлений ИНС является распознавание эмоций человека, в частности, на основе анализа речи (Speech Emotion Recognition, SER). Эта задача находится на стыке обработки естественного языка (Natural Language Processing, NLP) и акустического анализа, поскольку эмоции могут выражаться как вербально (слова и конструкции), так и невербально (интонация, тембр). Распознавание эмоций в речи является примером обучения с учителем (supervised learning) — метода машинного обучения, при котором модель обучается на размеченных данных, где каждому входному объекту (в данном случае, речевому сигналу) сопоставлен выходной сигнал (метка эмоции).

Технология распознавания эмоций в человеческой речи имеет потенциал применения в различных сферах. Она может использоваться для создания чат-ботов, способных лучше понимать эмоциональное состояние и намерения пользователя [1]. Другое применение — системы мониторинга эмоционального состояния в колл-центрах для определения настроения клиентов и повышения качества обслуживания [2]. Кроме того, анализ эмоций в речи может применяться в психиатрии для диагностики психических расстройств на ранней стадии [3]. В этих и многих других областях технология распознавания эмоций в речи открывает новые возможности и имеет высокую практическую ценность.

Традиционные подходы к распознаванию эмоций, основанные на экспертных правилах и классических методах машинного обучения, демонстрируют ограниченную эффективность из-за сложности и вариативности речевого сигнала, а также субъективности выражения эмоций. В то же время современные методы глубокого обучения, такие как сверточные и рекуррентные нейронные сети, показывают достойные результаты в задачах анализа речи и позволяют автоматически выявлять сложные закономерности в данных [4].

Важно отметить, что в рамках данной работы рассматривается частный случай задачи распознавания эмоций в речи, а именно задача, связанная с речью актеров. Актерская речь обладает рядом особенностей, отличающие ее от естественной речи. Эти и другие особенности будут рассмотрены в последующих разделах.

Актуальность темы обусловлена растущей потребностью в системах распознавания эмоций в речи и их высокой практической ценностью в различных областях применения.

Цель исследования: анализ влияния обучающей выборки и извлекаемых признаков на распознавание эмоций в человеческой речи с применением сверточных нейронных сетей.

Объект исследования: обучающие выборки, состоящие из речевых сигналов, содержащих эмоциональную информацию.

Предмет исследования: влияние состава и свойств обучающих выборок на обучение сверточных нейронных сетей.

Гипотеза исследования заключается в том, что структура, состав и свойства обучающей выборки оказывают значительное влияние на качество обучения нейросетевой модели.

Задачи исследования:

- провести анализ психологических и лингвистических аспектов выражения эмоций в речи, влияющих на формирование выборки данных и сформировать выборки с учетом данных аспектов;
- определение архитектуры сверточной нейронной сети для распознавания эмоций;
- провести обучение модели на подготовленных выборках;
- оценить эффективность обучения используемой модели на различных выборках;
- определить факторы, влияющие на качество обучения нейросетевой модели и сформулировать возможные способы повышения качества обучающих выборок.

Методы исследования:

- анализ научной литературы по проблеме распознавания эмоций, изучение современных подходов к анализу эмоций в речевом сигнале и формированию обучающих выборок для моделей глубокого обучения, содержащих речевые сигналы;
- методы анализа и обработки звуковых сигналов;
- методы глубокого обучения;
- методы количественного анализа результатов, включая вычисление метрик качества, анализ ошибок и визуализацию работы модели.

В данном исследовании применялась методология CRISP-DM (Cross-Industry Standard Process for Data Mining) — широко используемый стандартизованный итерационный процесс в области анализа данных, необходимый для структурирования и систематизации проектов. Она состоит из шести этапов:

- понимание бизнес-задачи: определение задач, требований, успешных критериев проекта;
- понимание данных: сбор данных, оценка качества и понимание их структуры;
- подготовка данных: подготовка данных для нейросетевой модели;
- моделирование: создание и обучение моделей машинного обучения;

- оценка: анализ результатов и проверка модели на соответствие поставленных целей и задач.
- внедрение:

Итеративный и циклический характер методологии, представленной на рис. 1, позволяет непрерывно совершенствовать решения и при необходимости возвращаться к предыдущим этапам для их пересмотра и внесения изменений.

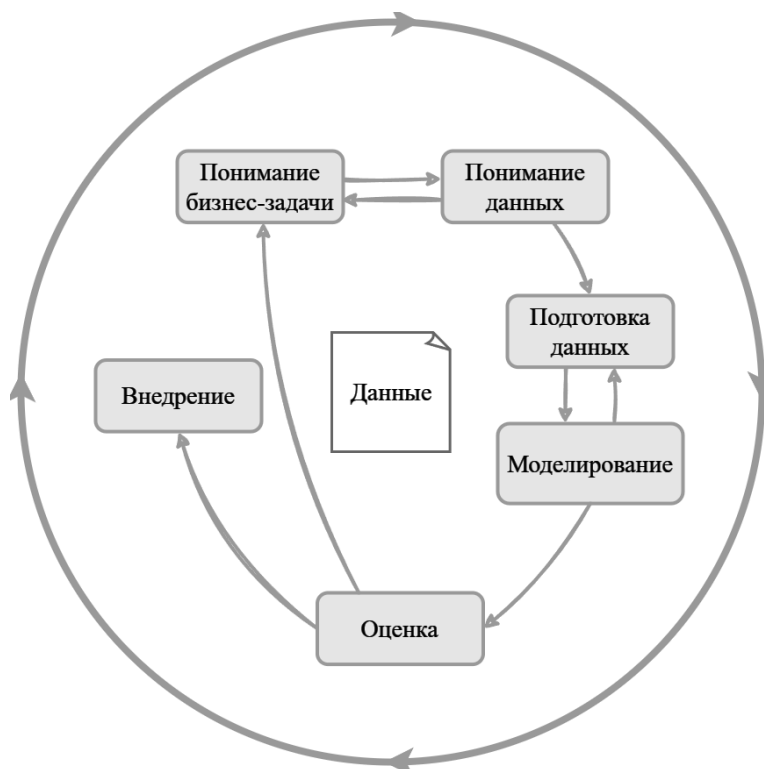


Рисунок 1. Методология CRISP-DM

Структура выпускной квалификационной работы включает в себя содержание работы, введение, четыре главы, заключение, список использованных источников и приложение.

Введение посвящено актуальности выбранной темы, определению целей и задач исследования, методов исследования, а также потенциал применения технологий распознавания эмоций в речи в различных областях.

В первой главе представлены теоретические основы для задачи распознавания эмоций в речи, а именно рассматриваются психологические и лингвистические аспекты выражения эмоций, обзор методов распознавания эмоций, анализ существующих наборов данных, подходы к подготовке данных для моделей глубокого обучения и метрики оценки качества моделей.

Во второй главе описывается методология исследования, включая выбор данных для обучения и тестирования модели, методы предобработки и подготовки данных, архитектуру предлагаемой нейросетевой модели, методы оценки качества модели и описание вычислительной платформы и программного инструментария, использованных для проведения экспериментов.

Третья глава содержит экспериментальное исследование, включая обучение и оценку эффективности модели, а также тестирование модели на различных наборах данных.

В четвертой главе проводится анализ результатов работы, включающий оценку качества распознавания эмоций на различных наборах данных и анализ факторов, влияющих на распознавание эмоций.

В заключении подводятся итоги исследования, обобщаются основные результаты и выводы, а также рассматриваются перспективы дальнейших исследований.

Список использованных источников включает источники, на которые опирается исследование. Приложение содержит код скриптов, которые представляют полный цикл решения задачи распознавания эмоций в человеческой речи.

1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

1.1 Психологические и лингвистические аспекты выражения эмоций в речи

Эмоции являются неотъемлемой частью человеческой коммуникации и оказывают существенное влияние на речевое поведение. Процесс выражения эмоций в речи представляет собой сложное взаимодействие психологических и лингвистических аспектов. В данном разделе рассматриваются психологические основы эмоций и особенности их проявления в речи с точки зрения лингвистики.

С психологической точки зрения, эмоции представляют собой комплексные реакции организма на внешние и внутренние стимулы, которые включают в себя физиологические изменения, когнитивные оценки ситуации и поведенческие проявления. Важно отметить, что эмоциональное состояние человека может проявляться по-разному в зависимости от ситуации, собеседника и социальной роли. Кроме того, существуют индивидуальные различия в выражении эмоций, связанные с личностными особенностями, социальной ролью, возрастом и полом [5].

Артикуляция является важным компонентом человеческой речи, играющим ключевую роль в выражении эмоций. Она представляет собой сложный процесс, который управляется нервной системой и осуществляется активными и пассивными органами речи. Активные органы (голосовые связки, язык, губы, вся нижняя челюсть и др.) производят звуки посредством движений, в то время как пассивные (зубы, твердое небо, вся верхняя челюсть и др.) формируют речевой тракт. Артикуляция звуков также зависит от положения и напряжения голосовых связок, а также от взаимодействия между ротовой и носовой полостями. Вибрация голосовых связок создает музыкальный тон, который используется при произнесении гласных и звонких согласных [6].

Важно различать естественные и искусственные эмоции в речи. Естественные эмоции отражают истинные чувства человека и являются спонтанными, выражаются в выделении гормонов и наблюдении физиологических изменений человека, в то время как искусственные эмоции, например, выражаемые актерами, являются результатом сознательного контроля, выражаются в ограниченном выделении гормонов и не имеют такой же физиологической основы [7].

Различия в языках мира играют значительную роль в выражении эмоций. Языки отличаются звуковыми системами и фонетическими особенностями, грамматическими и синтаксическими различиями, влияющие на структуру и порядок слов в предложении, что может повлиять на передачу эмоций. Культурный контекст также влияет на выражение эмоций. В некоторых культурах эмоциональная экспрессивность считается нормой, в то время как в других приветствуется сдержанность в выражении эмоций [8].

С лингвистической точки зрения, выражение эмоций в речи происходит на разных уровнях:

- *просодический уровень*: интонация, тембр, темп речи;
- *лексический уровень*: выбор эмоционально окрашенных слов;
- *синтаксический уровень*: структура предложений, паузы.

Исследования показывают, что каждая базовая эмоция имеет свои характерные акустические и просодические корреляты [9]. Например, радость часто ассоциируется с высоким тоном голоса, быстрым темпом речи и широким диапазоном частот основного тона, в то время как грусть характеризуется низким тоном, медленным темпом и малой вариативностью частоты основного тона.

Однако, несмотря на наличие общих закономерностей, выражение эмоций в речи отличается значительной вариативностью и зависит от межкультурных, контекстуальных и индивидуальных различий. Просодические характеристики, такие как интенсивность, темп и высота тона речи могут отражать культурные различия в выражении эмоций [10].

В данной работе исследование ограничивается взрослыми людьми, а именно актерами, использующими английскую речь. Языки тонического ударения, в которых смысловое значение слов определяется изменением тона голоса, не рассматриваются. Кроме того, в работе анализируются только утвердительные по цели высказывания предложения, которые сами по себе не несут эмоциональной окраски. Исследование фокусируется исключительно на голосовых проявлениях эмоций, встречающихся в речи.

Таким образом, разработка систем автоматического распознавания эмоций в речи требует комплексного подхода, учитывающего не только общие закономерности выражения эмоций, но и межкультурные, контекстуальные и индивидуальные различия. Интеграция знаний из области психологии, лингвистики и межкультурной коммуникации является необходимым условием для создания эффективных и надежных моделей распознавания эмоций, применимых в реальных условиях.

1.2 Обзор методов распознавания эмоций

Методы распознавания эмоций в речи можно условно разделить на традиционные подходы и подходы, основанные на глубоких нейронных сетях. Сравнение этапов традиционного машинного обучения и глубокого обучения в распознавании эмоций представлены на рис. 1.1.

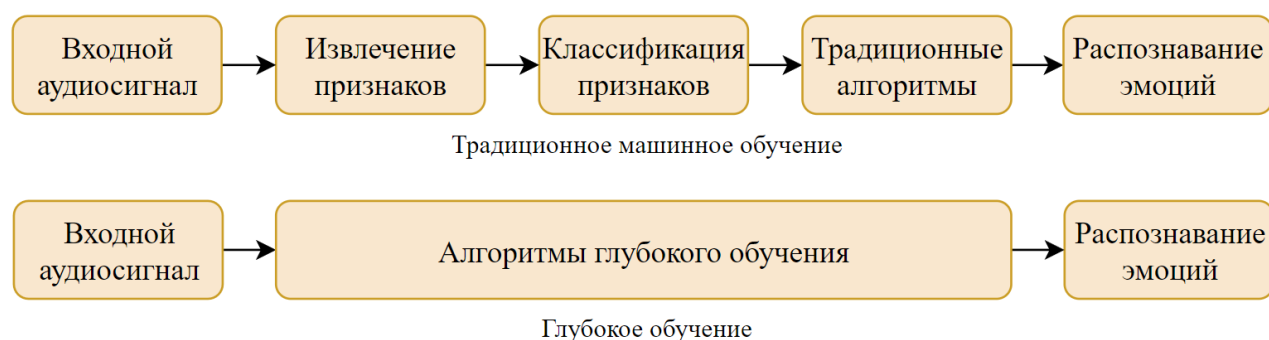


Рисунок 1.1 Этапы традиционного машинного обучения и глубокого обучения

Традиционные методы, активно развивавшиеся в 1990-2000х годах, основаны на экспертных правилах и классических алгоритмах машинного обучения, таких как скрытые марковские модели (HMM), метод опорных векторов (SVM), искусственные нейронные сети (ANN), метод ближайших соседей (KNN) и др. Эти подходы предполагают ручное извлечение признаков из речевого сигнала, например, статистики частоты основного тона или энергии, с последующей классификацией эмоций на основе извлеченных признаков. Несмотря на определенные успехи, традиционные методы демонстрируют ограниченную эффективность из-за сложности и вариативности эмоциональной речи, а также трудоемкости разработки признаков вручную [4].

В последнее десятилетие значительный прогресс в области распознавания эмоций в речи был достигнут благодаря развитию методов глубокого обучения и применению глубоких нейронных сетей. Термин «глубокое обучение» происходит от использования нейронных сетей с большим количеством скрытых слоев, что позволяет моделировать сложные нелинейные зависимости в данных. Скрытые слои последовательно преобразуют входные данные, извлекая из них иерархические представления различного уровня абстракции. Такие архитектуры, как сверточные нейронные сети (CNN), рекуррентные нейронные сети (RNN) и трансформеры позволяют автоматически выявлять сложные закономерности в речевых данных и строить высокоуровневые представления, и как следствие, эффективно применяться в рассматриваемой задаче [4][11].

Ключевым преимуществом глубоких нейронных сетей является их способность самостоятельно обучаться информативным признакам из необработанных данных, например, необработанные речевые сигналы или спектрограммы. Это существенно упрощает процесс разработки систем распознавания эмоций и обеспечивает более высокую точность по сравнению с традиционными подходами [4].

CNN особенно эффективны для извлечения локальных спектрально-временных признаков из речевого сигнала. Они состоят из нескольких слоев, включая сверточные слои (Convolution Layer), слои объединения (Pooling Layer) и полносвязные слои (Dense Layer). Сверточные слои применяют набор фильтров к входным данным, выявляя локальные признаки и формируя карты признаков. Слои объединения уменьшают размерность карт признаков, сохраняя наиболее важную информацию. Полносвязные слои в конце сети объединяют извлеченные признаки и выполняют классификацию эмоций. CNN способны

автоматически обучаться иерархическим представлениям признаков, от низкоуровневых спектральных характеристик до высокоуровневых эмоциональных репрезентаций [4].

RNN, такие как долгая краткосрочная память (Long Short-Term Memory, LSTM) и управляемые рекуррентные блоки (Gated Recurrent Unit, GRU), специально разработаны для обработки последовательных данных и моделирования долгосрочных зависимостей. Они содержат рекуррентные связи, позволяющие сохранять информацию о предыдущих состояниях сети и использовать ее для обработки текущего входа. LSTM и GRU имеют специальные механизмы гейтов, которые регулируют поток информации и помогают избежать проблемы исчезающего и взрывающегося градиента [12].

Трансформеры — современная архитектура глубокого обучения, основанная на механизме внимания (self-attention). В отличие от CNN и RNN, трансформеры не опираются на фиксированную последовательность входных данных и могут моделировать зависимости между любыми парами элементов последовательности. Механизм внимания позволяет модели динамически определять важность различных участков речевого сигнала для распознавания эмоций [11].

Помимо использования отдельных архитектур, исследователи также разрабатывают гибридные модели, объединяющие преимущества различных подходов. Например, была эффективно использована комбинация CNN и LSTM (CNN-LSTM), позволяющая извлекать локальные спектрально-временные признаки с помощью сверточных слоев и моделировать их временную динамику с помощью рекуррентных слоев [13]. Другим примером является использование трансформеров совместно с CNN [14].

В данной работе используется только сверточная нейронная сеть. Это обусловлено тем, что ее часто используют в задаче распознавания эмоций в речи, а также ее использование достаточно для решения задачи.

Таким образом, область распознавания эмоций в речи активно развивается, и современные методы глубокого обучения открывают новые возможности для создания более эффективных и надежных систем.

1.3 Наборы данных для распознавания эмоций

Наборы данных играют ключевую роль в разработке и оценке систем распознавания эмоций в речи. Они служат основой для обучения, настройки и тестирования моделей машинного обучения и глубоких нейронных сетей. От качества и разнообразия наборов данных во многом зависит, насколько хорошо полученные модели будут работать в реальных условиях [4].

Обычно набор данных для распознавания эмоций состоит из аудиозаписей эмоциональной речи, размеченных по определенной системе эмоциональных

меток. Метки обозначают основные категории эмоций (радость, грусть, гнев, страх и т. д.).

Наборы данных могут содержать речевой материал разного типа, полученный различными способами [4]:

- *студийная запись актерской речи*: профессиональные актеры разыгрывают эмоциональные сценарии в контролируемых условиях звукозаписывающей студии;
- *естественная спонтанная речь*: запись реальных диалогов и взаимодействий людей в повседневных или специфических ситуациях (колл-центры, медицинские консультации и т. д.);
- *индуцированная эмоциональная речь*: у участников записи вызываются определенные эмоции с помощью специальных стимулов или сценариев;
- *сгенерированная эмоциональная речь*: использование искусственных нейронных сетей для генерации эмоций, тем самым увеличивая данные и создавая образцы под свои требования.

Важно, чтобы наборы данных были достаточно большими по объему (количество аудиозаписей и общая длительность), разнообразными по дикторам (пол, возраст, язык, акцент), имели хорошее качество записи (без лишних шумов и искажений) и содержали одинаковое количество примеров каждого класса эмоций. Чем больше и разнообразнее набор данных, тем более устойчивые и эффективные модели можно на нем обучить [15]. Кроме того, важно обеспечить фонетическую сбалансированность данных, необходимую для минимизации влияния индивидуальных особенностей голоса и произношения на распознавание эмоций.

Помимо аудиоданных, наборы могут включать дополнительную информацию, такую как расшифровки речи, языковые особенности, электромиографию (ЭМГ), электроэнцефалографию (ЭЭГ), видеозаписи лица и жестов говорящего [16]. Эти мультимодальные данные позволяют изучать и моделировать связи между различными способами выражения эмоций и создавать более совершенные системы распознавания.

Таким образом, качественные и разнообразные наборы данных — ключевой фактор успеха в развитии систем распознавания эмоций в речи.

1.4 Подготовка данных для моделей глубокого обучения

Звуковые сигналы обычно представлены в виде аудиозаписей в цифровом формате, полученные в результате аналого-цифрового преобразования некоторого аналогового сигнала. В дальнейшем эти аудиозаписи требуют предварительной обработки перед подачей на вход модели искусственной нейронной сети. Для этого применяются методы извлечения признаков, разделения данных на выборки, нормализации и аугментации данных.

Признаки — это характеристики, извлекаемые из речевого сигнала [17]. Каждый признак характеризует конкретное проявление эмоций и речи. Их комбинирование позволяет модели эффективно различать основные эмоции. Признаки могут быть чистыми, однозначно определяющие какой-то параметр, и смешанными. В природе все признаки смешанные, однако в результате комбинации смешанных оценок может выйти чистый признак. Этот феномен известен как синергетический эффект. Он предполагает, что использование комбинации нескольких подходов дает более лучший результат, чем применение каждого подхода по отдельности.

Существует множество методов обработки сигналов для извлечения признаков, среди которых можно выделить методы разделения сигнал/шум, голос/шум, голос/голос, методы определения речи в голосе и музыки в голосе. Кроме того, существуют методы определения эмоций, такие как анализ просодических характеристик речи и спектральных особенностей голоса.

Признаки можно поделить на три основные категории [4]:

- *просодические*: ритм, интонацию и другие свойства речи на уровне предложений, слов, слогов и выражений;
- *спектральные*: распределение энергии сигнала по частотам;
- *качества голоса*: характеристики голоса, связанные с работой голосовых связок и артикуляционного аппарата;

В данной работе использовались только просодические и спектральные признаки: Zero Crossing Rate (ZCR), Root Mean Square Energy (RMSE) и Mel-Frequency Cepstral Coefficients (MFCC) [17].

ZCR — просодический признак, представляющий собой частоту изменения знака амплитуды сигнала. Она отражает темп речи. ZCR вычисляется по формуле (1.1):

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} \frac{|sgn(x[n]) - sgn(x[n-1])|}{2} \quad (1.1)$$

где $x[n]$ — амплитуда сигнала в момент времени n ;

N — длина сигнала;

sgn — функция знака.

RMSE — просодический признак, отражающий интенсивность речевого сигнала. Вычисляется по формуле (1.2):

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N x[n]^2} \quad (1.2)$$

где $x[n]$ — амплитуда сигнала в момент времени n ;

N — длина сигнала.

Представленные просодические признаки зависят от произнесенных слов, поскольку акустическая структура разных слов и звуков различается. Например, слова с большим количеством согласных обычно имеют более высокое значение ZCR.

MFCC — спектральный признак, представляющий собой набор коэффициентов, описывающих огибающую спектра сигнала в мел-шкале частот, которая учитывает особенности восприятия звука человеческим ухом и отражает психоэмоциональное состояние человека. Набор коэффициентов получают путем применения дискретного косинусного преобразования (DCT) к логарифму мел-спектра сигнала. MFCC является уникальным для каждого человека и отражает изменение тона речи. Пример процесса вычисления MFCC:

1. Сначала происходит преобразование аудиофайла в массив со значениями амплитуд $x[n]$ в момент времени n , где $n = 0, 1, \dots, N - 1$.
2. Применение фильтра предварительного усиления к сигналу для усиления высоких частот, чтобы сбалансировать частотный спектр, избежать проблем в вычислениях во время преобразования Фурье, улучшить отношение сигнал/шум (SNR):

$$s[n] = x[n] - ax[n-1] \quad (1.3)$$

где a — коэффициент фильтра усиления (обычно 0.95).

3. Сигнал разбивается на фреймы, обычно с длительностью от 20 до 40 мс с 50% перекрытием соседних фреймов.
4. На каждом фрейме применяется функция $w[n]$ (например, окно Хэмминга):

$$s_w[n] = s[n] \cdot w[n], \quad w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad (1.4)$$

где $0 \leq n \leq N - 1$, N — длина окна.

5. Для каждого фрейма $s_w[n]$ выполняется N -точечное быстрое преобразование Фурье (FFT) для получения частотного спектра.

$$X[k] = \sum_{n=0}^{N-1} s_w[n] \cdot e^{-\frac{j2\pi kn}{N}} \quad (1.5)$$

Затем вычисляется спектр мощности (периодограмма) по формуле (1.6).

$$P[k] = \frac{|X[k]|^2}{N} \quad (1.6)$$

где k — индекс частоты;

N — длина фрейма.

6. Спектр мощности преобразуется в мел-шкалу с использованием банка треугольных фильтров (обычно 40 фильтров). Мел-шкала имитирует нелинейное восприятие звука человеческим ухом. Она более различима на низких частотах и менее различима на более высоких. Энергия в каждой полосе фильтра рассчитывается по формуле (1.7).

$$M[m] = \sum_{k=f_{start}}^{f_{end}} P[k] \cdot H_m[k] \quad (1.7)$$

где $M[m]$ — энергия на выходе m -го фильтра;

$H_m[k]$ — треугольные фильтры;

f_{start} и f_{end} — границы частот фильтра.

Зависимость между герцами f и мел-шкалой m :

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right), f = 700 \left(10^{\frac{m}{2595}} - 1 \right) \quad (1.8)$$

7. Применяется логарифм к выходным значениям фильтров.

$$\log_{10} M[m] \quad (1.9)$$

8. Применяется DCT к логарифмированным значениям для декорреляции признаков и уменьшения размерности, получая MFCC.

$$MFCC[n] = \sum_{m=0}^{M-1} \log_{10} M[m] \cdot \cos \left(\frac{\pi n(2m+1)}{2M} \right) \quad (1.10)$$

где $MFCC[n]$ – n -ый кепстральный коэффициент;

M – количество фильтров;

$n = 0, 1, \dots, K-1$; K – количество коэффициентов (обычно 13).

Стоит отметить, что при вычислении признаков используются разные параметры, и в зависимости от этих параметров получаются разные признаки. Поэтому в зависимости от задачи следует тщательно их подбирать для получения лучших результатов. Подбор параметров — сложная задача, которой посвящено много исследовательских работ. В данной работе были выбраны параметры, которые встречались в документации librosa версии 0.10.1 [18].

После извлечения признаков они объединяются в единый вектор признаков для каждого аудиофрагмента. Далее для приведения признаков к одному масштабу и улучшения сходимости модели при обучении используется нормализация данных. Существуют различные методы нормализации, такие как min-max нормализация, z-нормализация и др. В данной работе используется метод стандартизации StandardScaler, который центрирует данные и масштабирует их таким образом, чтобы они имели нулевое среднее значение и единичную дисперсию. Сначала вычисляется среднее значение признака по формуле (1.11).

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (1.11)$$

где N – количество значений признака;

x_i – значение признака для i -го наблюдения.

Далее по формуле (1.12) вычисляется дисперсия признака.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (1.12)$$

Наконец, получаем нормализованный признак по формуле (1.13).

$$x_{norm} = \frac{x - \mu}{\sigma} \quad (1.13)$$

где x – исходное значение признака;

μ – среднее значение признака;

σ – дисперсия признака.

Поскольку задача распознавания эмоций является задачей классификации, необходимо преобразовать категориальные метки эмоций в числовой формат,

понятный модели. Для этого используется OneHotEncoder, который преобразует категориальные метки в бинарные векторы.

Обычно данные разделяют на несколько выборок, каждая из которых нужна для разных целей. Наиболее распространенным является разделение на обучающую (training set), валидационную (validation set) и тестовую (test set) выборки. Как правило, данные разделяют в соотношении 80/20 или 70/30 (обучающая/тестовая) или 60/20/20 (обучающая/валидационная/тестовая). Модель обучается на обучающей выборке, ее параметры корректируются для минимизации ошибок на этой выборке. Валидационная выборка используется для настройки гиперпараметров модели и оценки ее способности к обобщению на новых данных в процессе обучения. Тестовая выборка, в свою очередь, используется только один раз после завершения обучения для финальной оценки качества модели.

Важно обеспечить сбалансированное представление данных в каждой выборке, особенно если речь идет о классификации. Например, если мы обучаем модель распознавать эмоции, и в обучающей выборке будет преобладать радость, модель покажет низкую точность на других эмоциях. Для решения этой проблемы применяют стратегии разделения данных, например, стратифицированная выборка, которая гарантирует пропорциональное представление каждой эмоции в каждой из выборок.

Аугментация данных нужна для увеличения объема и повышения вариативности обучающей выборки для повышения обобщающей способности модели, особенно при ограниченном объеме обучающих данных. В контексте аудиоданных и задачи распознавания эмоций, аугментация позволяет имитировать разнообразие акустических условий и вариативность проявления эмоций в речи. В данной работе применялись следующие методы аугментации:

- *добавление случайного шума*: имитирует фоновый шум, который может присутствовать в реальных условиях записи речи;
- *временной сдвиг*: небольшой сдвиг сигнала по времени позволяет модели лучше адаптироваться к различным темпам речи и незначительным вариациям в произношении;
- *изменение высоты тона*: имитирует различия в высоте голоса у разных людей, что способствует более робастному распознаванию эмоций независимо от индивидуальных особенностей говорящего;
- *комбинация методов*: совместное применение нескольких методов аугментации позволяет создать еще большее разнообразие обучающих примеров и сделать модель более устойчивой к вариациям во входных данных.

Исходная аудиозапись и примененные к ней методы аугментации изображены на рис. 1.2–1.6.

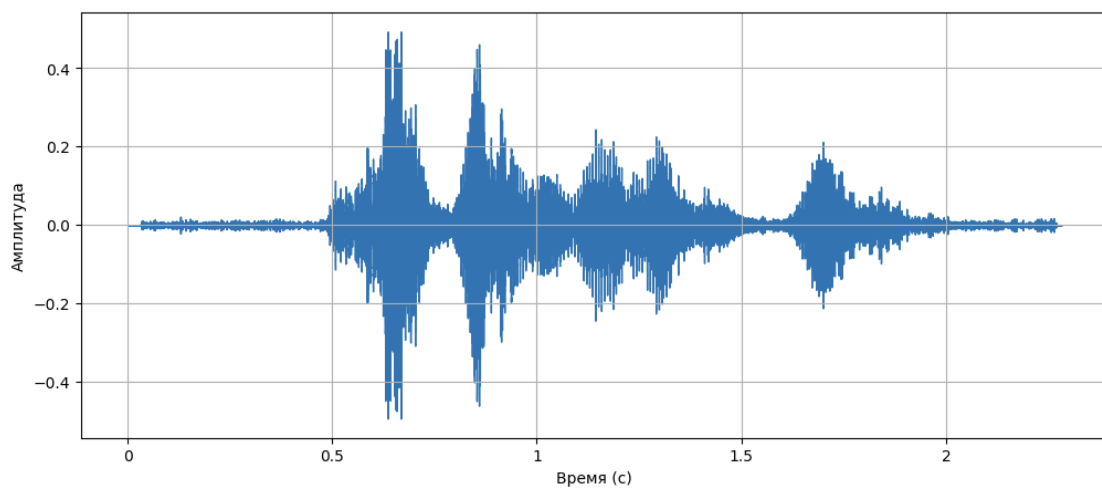


Рисунок 1.2 Исходная аудиозапись

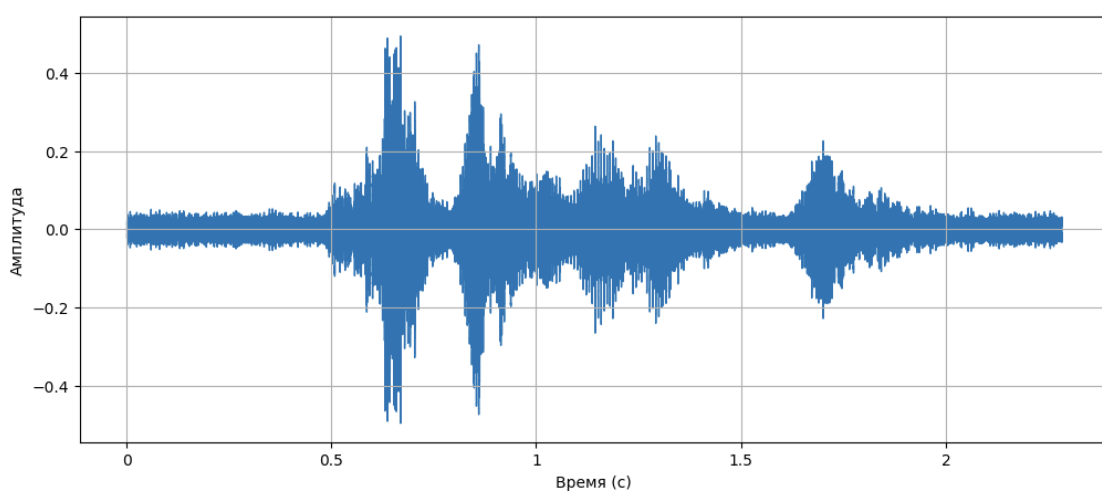


Рисунок 1.3 Исходная аудиозапись с добавлением случайного шума

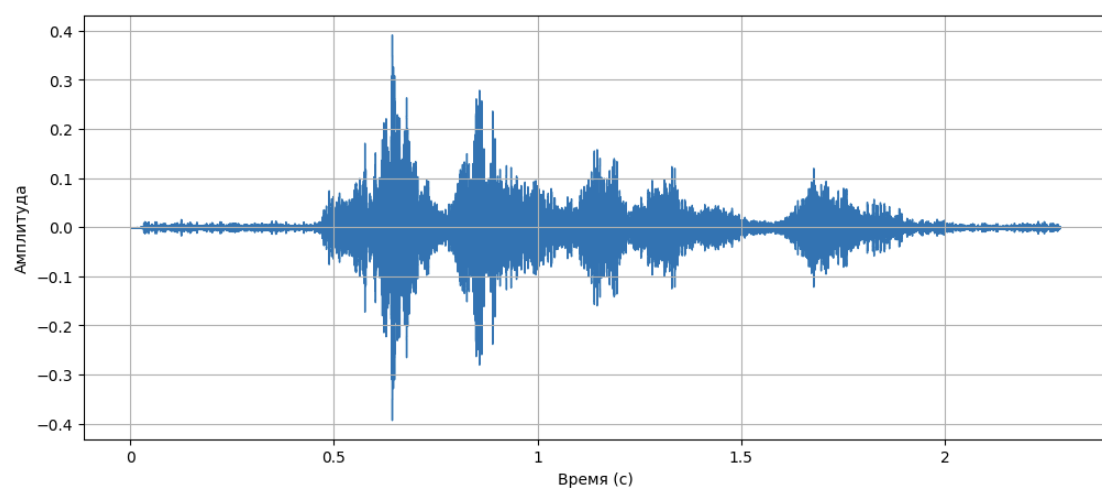


Рисунок 1.5 Исходная аудиозапись с изменением высоты тона

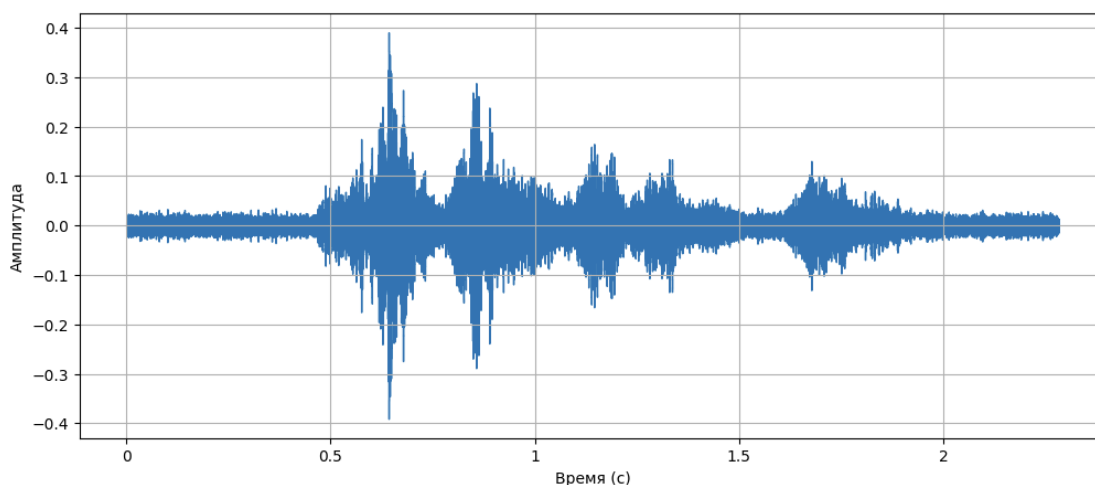


Рисунок 1.6 Исходная аудиозапись с добавлением шума и изменением высоты тона

В результате подготовки данных для модели нейронной сети получают обучающую, состоящую из признаков, и валидационную выборки, готовые для подачи на вход нейросетевой модели. Извлечение признаков — основа для формирования информативного представления эмоций, а проведенные преобразования и аугментация данных позволили увеличить размер обучающей выборки и повысить устойчивость модели к вариациям в исходных аудиоданных.

1.5 Метрики оценки качества моделей

Для оценки эффективности моделей распознавания эмоций и их сравнения с другими подходами используется комплекс метрик, позволяющих количественно измерить производительность. Выбор конкретных метрик зависит от специфики задачи, типа используемых эмоциональных меток и характера данных.

Прежде чем перейти к описанию метрик, определим ключевые показатели, используемые в расчете метрик в задачи классификации [19]:

- *TP (True Positive)* — истинно положительные результаты: количество образцов, где модель корректно определила наличие определенной эмоции.
- *TN (True Negative)* — истинно отрицательные результаты: количество образцов, где модель корректно определила отсутствие определенной эмоции.
- *FP (False Positive)* — ложно положительные результаты: количество образцов, где модель ошибочно определила наличие определенной эмоции.

- *FN (False Negative)* — ложно отрицательные результаты: количество образцов, где модель ошибочно определила отсутствие определенной эмоции.

В случае классификации эмоций с дискретными метками применяются следующие метрики [19]: Accuracy, Precision, Recall, F1-score, Confusion Matrix.

Accuracy (точность) отражает общую долю правильных предсказаний модели относительно общего числа образцов. Эта метрика позволяет оценить общую эффективность модели в правильной классификации образцов, вычисляется по формуле (1.14).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1.14)$$

Precision (точность) показывает, какая доля образцов, отнесенных моделью к данному классу, действительно принадлежит этому классу. Вычисляется по формуле (1.15).

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (1.15)$$

где TP_i – истинно положительные результаты для i -го класса;

FP_i – ложно положительные результаты для i -го класса.

Recall (полнота) показывает, какую долю образцов данного класса модель смогла обнаружить. Вычисляется по формуле (1.16).

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (1.16)$$

где TP_i – истинно положительные результаты для i -го класса;

FN_i – ложно отрицательные результаты для i -го класса.

F1-score (F1-мера) представляет собой гармоническое среднее между точностью (Precision) и полнотой (Recall), учитывающее баланс между этими двумя показателями. F1-score вычисляется как среднее гармоническое значение Precision и Recall по всем классам эмоций и представляется в виде формулы (1.17).

$$F1_i = \frac{Precision_i \cdot Recall_i}{Precision_i + Recall_i} \quad (1.17)$$

Confusion Matrix (матрица ошибок) — таблица, которая визуализирует результаты классификации, показывая количество правильно и неправильно классифицированных образцов для каждого класса. На главной диагонали матрицы расположены правильно классифицированные образцы, в то время как вне диагонали находятся образцы, ошибочно отнесенные к другим классам. Анализ матрицы ошибок позволяет выявить типичные ошибки модели и понять, какие классы эмоций наиболее часто путаются между собой.

Использование комплекса метрик, таких как Accuracy, Precision, Recall, F1-score и Confusion Matrix, позволяет всесторонне оценить качество модели классификации эмоций, учитывая общую точность, эффективность распознавания отдельных классов и типичные ошибки предсказания. Данные метрики были выбраны как наиболее распространенные и достаточно информативные для оценки качества классификации. Важно отметить, что

существуют и другие метрики, например, ROC (Receiver Operating Characteristic), AUC (Area Under the Curve) и др. [19], однако в рамках данной работы их применение не рассматривается.

Таким образом, метрики оценки качества играют важную роль в процессе разработки и совершенствования моделей распознавания эмоций в речи. Они позволяют количественно измерять эффективность различных подходов, выявлять их сильные и слабые стороны, а также определять дальнейшие исследования в этой области.

2 МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

2.1 Выбор данных для обучения и тестирования модели

В данной квалификационной работе для решения задачи распознавания эмоций в речи были использованы четыре общедоступных набора данных: CREMA-D, RAVDESS, SAVEE и TESS. Эти наборы данных широко применяются в исследованиях по автоматическому распознаванию эмоций [20-22] и содержат аудиофайлы с актерской речью с выражением различных эмоциональных состояний.

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) [23] содержит видео и аудиозаписи речи актеров, которые произносят 12 различных предложений. Набор данных включает 7442 аудиофайла от 91 актера (48 мужчин и 43 женщины) с различными расовыми принадлежностями и этническими группами (афроамериканцы, азиаты, европеоиды, латиноамериканцы и неуказанные) в возрасте от 20 до 74 лет. Эмоции представлены шестью основными классами: счастье, печаль, гнев, страх, отвращение и нейтральное состояние. Каждая аудиозапись была оценена по эмоциональной интенсивности (низкая, средняя, высокая, неуказанная). Оценка данных была получена от 2443 респондентов путем привлечения широкого круга пользователей к оценке. Примеры используемых фраз: «Don't forget a jacket», «It's eleven o'clock», «I would like a new alarm clock», «That is exactly what happened» и «I'm on my way to the meeting».

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [24] содержит аудиозаписи и видеозаписи речи актеров, выражающих различные эмоции в речи и пении. Набор данных включает 1440 файлов с речью от 24 профессиональных актеров (12 женщин и 12 мужчин), которые произносят два лексически согласованных утверждения с нейтральным североамериканским акцентом. Эмоции представлены восемью категориями: нейтральное состояние, спокойствие, счастье, печаль, гнев, страх, отвращение и удивление, а также используются два уровня эмоциональной интенсивности (нормальный и сильный). Актеры произносят два предложения: «Kids are talking by the door» и «Dogs are sitting by the door».

SAVEE (Surrey Audio-Visual Expressed Emotion) [25] — это набор аудиовизуальных данных, содержащий записи 4 британских актеров мужчин в возрасте от 27 до 31 года, произносящих 15 фонетически сбалансированных предложений с 7 различными эмоциями: гнев, отвращение, страх, счастье, нейтральное состояние, печаль и удивление. Всего в наборе данных содержится 480 аудиозаписей, по 120 на каждого актера. Примеры предложений: «She had your dark suit in greasy wash water all year» и «Don't ask me to carry an oily rag like that».

TESS (Toronto Emotional Speech Set) [26] содержит 2800 аудиозаписей от двух актрис из Канады в возрасте 26 и 64 лет, произносящих 200 фраз по шаблону

«Say the word _», представленных семью эмоциями: гнев, отвращение, страх, счастье, нейтральное состояние, печаль и приятное удивление. Примеры фраз: «Say the word bite», «Say the word came», «Say the word mouse».

Все рассматриваемые наборы данных содержат размеченную информацию в названиях аудиофайлов. Такой подход к хранению метаданных обеспечивает эффективность и удобство работы с данными. Информацию о говорящем, эмоции, уровне интенсивности и других параметрах извлекаются из имени файла, что упрощает процесс обработки данных.

В данной работе рассматриваются пять сбалансированных классов эмоций: Anger, Disgust, Fear, Happy и Sad. Количество записей в наборах данных, классифицированных по классам эмоций, представлены в таблице 2.1. Класс Neutral имеет разное количество записей в наборах данных, классы Surprise и Calm содержат слишком малое количество записей, а в некоторых наборах эти классы отсутствуют. Далее эти наборы данных объединяются в один, получая один набор данных, состоящий из 9615 записей.

Таблица 2.1 – Количество записей в наборах данных по классам эмоций

Наборы данных	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise	Calm	Итого
CREMA-D	1271	1271	1271	1271	1087	1271	0	0	7442
RAVDESS	192	192	192	192	96	192	192	192	1440
SAVEE	60	60	60	60	120	60	60	0	480
TESS	400	400	400	400	400	400	400	0	2800
Объединенный набор данных	1923	1923	1923	1923	1703	1923	652	192	12162

Несмотря на то, что эти наборы данных могут казаться взаимодополняемыми благодаря разнообразию эмоций и вариативности дикторов, их объединение для обучения модели распознавания эмоций может оказаться неэффективным по ряду причин:

- наблюдается несогласованность в представлении эмоций между наборами данных. CREMA-D включает в себя 6 эмоций, RAVDESS — 8, SAVEE и TESS — 7. При этом некоторые эмоции, например, «удивление» в RAVDESS и «приятное удивление» в TESS имеют схожие названия, но при этом отражают разные оттенки эмоции. Даже если рассмотреть одинаковые эмоции, то они выражаются по-разному. Объединения таких данных может привести к размыванию границ между классами и затруднит обучение;
- имеется значительное расхождение в количестве записей между наборами данных. CREMA-D содержит на каждую эмоцию по 1271 записей, а SAVEE на тех же эмоциях имеет всего по 60 записей, RAVDESS — по 192 записи. Обучение этих данных вместе может повлечь за собой то, что

модель будет непропорционально сильно ориентироваться на особенности CREMA-D, что негативно скажется на ее обобщающие способности;

- наборы данных имеют разную структуру высказываний, лексический состав и стиль речи. В таком случае обучение на объединенных данных может привести к смещению модели в сторону запоминания специфических фраз, а не обобщению акустических признаков, характерных для конкретных эмоций.
- наборы данных записаны с участием разного количества дикторов, отличающихся по полу, возрасту и профессиональным качествам. Если рассмотреть пол, то CREMA-D включает в себя записи 48 мужчин и 43 женщины в возрасте от 20 до 74 лет, в то время как SAVEE ограничен записями 4 мужчин в возрасте от 27 до 31, а TESS — 2 женщинами в возрасте 26 и 64 года. Объединение таких данных может привести к тому, что модель будет неравномерно чувствительна к эмоциям, выраженным разными демографическими группами;
- эмоции в этих наборах искусственные и сильно преувеличенные. Актеры стремятся к передаче максимально четкой и однозначной эмоции, что приводит к некоторой «театральности», когда в реальной жизни люди выражают эмоции многогранно и на гормональном уровне.

Таким образом, объединение рассматриваемых наборов данных представляется нецелесообразным. Однако, в некоторых работах эти данные объединяются для обучения и утверждается, что модель успешно распознает эмоции [20][22]. Необходимо провести анализ распознавания эмоций моделью в зависимости от наборов данных и подтвердить предположения.

2.2 Подготовка данных

Подготовка данных состоит из четырех важных этапов: извлечение признаков из аудиофайлов, аугментация, нормализация данных и разделение данных на выборки.

Для извлечения признаков из аудиоданных были использованы методы ZCR, RMSE, MFCC, реализованные в библиотеке librosa, со следующими параметрами:

- *frame_length=2048*: определяет длину окна в фреймах.
- *hop_length=512*: определяет шаг, с которым окно перемещается по аудиосигналу в фреймах.
- *n_mfcc=13*: указывает количество коэффициентов MFCC, которые нужно извлечь из каждого окна.
- *n_fft=frame_length*: на каждое окно производится одно преобразование Фурье.
- *sr=22050*: частота дискретизации аудиофайла, измеряемая в герцах, влияет на диапазон частот, представленных в MFCC.

Размер окна 2048 образцов обеспечивает достаточное разрешение по частоте для анализа спектральных характеристик аудио, в то время как шаг 512 образцов позволяет отслеживать быстрые изменения сигнала во времени. Извлечение MFCC коэффициентов в размере 13 является распространенной практикой и позволяет охватить основные спектральные характеристики звука. Выбранная частота дискретизации 22050 Гц является стандартной для аудиофайлов и обеспечивает хорошее соотношение качества и размера данных.

Для увеличения разнообразия обучающих данных и повышения устойчивости модели к шумам были применены методы аугментации данных. Каждый аудиофрагмент подвергался следующим преобразованиям:

- Добавление случайного шума к исходному сигналу;
- Сдвиг сигнала на случайное количество отсчетов вперед или назад;
- Изменение высоты тона сигнала;
- Комбинация изменения высоты тона и добавления шума.

Таким образом, для каждого исходного аудиофрагмента было получено четыре новых примера, что позволило увеличить размер обучающей выборки и улучшить обобщающую способность модели.

Извлечение признаков было выполнено параллельно на 12 логических процессорах с использованием библиотеки `joblib`, что намного увеличило скорость извлечения. Полученные признаки были сохранены в `DataFrame` вместе с соответствующими метками эмоций и путями к исходным аудиофайлам.

Для приведения признаков к одному масштабу и улучшения сходимости модели при обучении используется нормализация данных. В данном случае используется метод стандартизации `StandardScaler`, а также были закодированы целевые метки эмоций с помощью метода `OneHotEncoder`.

Для обеспечения репрезентативности и надежности оценки модели было реализовано стратифицированное разделение данных на обучающую и валидационную выборки. Код выполняет разделение для каждой комбинации эмоции и набора данных отдельно, выделяя фиксированное количество примеров в валидационную выборку в размере 20% от всех данных. Такой подход гарантирует, что валидационная выборка будет содержать репрезентативное и равное количество примеров для каждой эмоции из каждого набора данных, что позволяет более точно оценить обобщающую способность модели на новых, ранее не встречавшихся данных.

Обученные `encoder` меток и `scaler` признаков были сохранены в файлы для дальнейшего использования при тестировании модели. Пути к аудиофайлам и соответствующие метки эмоций для обучающей и тестовой выборок были сохранены в отдельные текстовые файлы.

В результате предобработки и подготовки данных были получены обучающая и валидационная выборки, готовые для подачи на вход нейросетевой модели. Проведенные преобразования и аугментация данных позволили увеличить размер обучающей выборки и повысить устойчивость модели к вариациям в исходных аудиоданных.

2.3 Предлагаемая архитектура нейросетевой модели

Для решения задачи распознавания эмоций в речи в данной работе предлагается использование архитектуры сверточной нейронной сети (CNN), представленной на рис. 2.1. Подробнее архитектура будет показана в разделе 1 главы 3. Модель построена с использованием библиотеки Keras и представляет собой последовательность слоев, каждый из которых выполняет определенные функции в процессе обработки входных данных.



Рисунок 2.1 Предлагаемая архитектура нейросетевой модели

Первые слои модели являются сверточными (Conv1D) и применяют операцию свертки к входным данным. Эти слои действуют как локальные фильтры, извлекая признаки из речевого сигнала. Каждый фильтр обучается распознавать определенные характеристики сигнала, в результате чего получаются карты признаков, представляющие собой локальные представления входного сигнала. Все сверточные слои используют функцию активации ReLU (Rectified Linear Unit), которая добавляет нелинейность в процесс обработки данных, что позволяет модели обучаться более сложным и абстрактным признакам. ReLU определяется по формуле (2.1):

$$ReLU(x) = \max(0, x) \quad (2.1)$$

где x – входное значение нейрона.

После сверточных слоев следуют слои пакетной нормализации (BatchNormalization), которые нормализуют активации предыдущего слоя. Пакетная нормализация помогает стабилизировать процесс обучения, ускорить сходимость и снизить чувствительность модели к выбору гиперпараметров. Она приводит активации к нулевому среднему значению и единичной дисперсии, что облегчает обучение последующих слоев.

Затем в модели используются слои объединения (MaxPool1D), которые уменьшают размерность карт признаков. Эти слои выбирают максимальное значение из локальных областей карты признаков, что позволяет сохранить наиболее важные признаки и уменьшить вычислительную сложность модели.

Для регуляризации модели и предотвращения переобучения применяются слои прореживания (Dropout). Они случайным образом отключают часть нейронов во время обучения, что заставляет модель обучаться более устойчивым и обобщенным признакам.

После последовательности сверточных, нормализационных и объединяющих слоев, карты признаков преобразуются в одномерный вектор с помощью слоя Flatten. Этот слой выравнивает многомерные карты признаков в одномерный вектор, который затем подается на вход полносвязным слоям (Dense).

Полносвязные слои выполняют роль классификатора. Они принимают на вход признаки, извлеченные сверточными слоями, и обучаются распознавать эмоции на основе этих признаков. Первый полносвязный слой имеет определенное количество нейронов и использует функцию активации ReLU. Второй полносвязный слой имеет количество нейронов, равное числу распознаваемых эмоций (num_emotions), и использует функцию активации softmax. Softmax преобразует выходы нейронов в вероятности принадлежности к каждому классу эмоций, так что сумма всех вероятностей равна 1. Математически функция softmax определяется следующим образом:

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (2.2)$$

где x_i – выход i -го нейрона;

j – индекс, проходящий по всем нейронам выходного слоя.

Процесс обучения модели заключается в настройке ее параметров (весов и смещений) таким образом, чтобы минимизировать ошибку между предсказанными и истинными метками эмоций. Для этого используется оптимизатор Adam (Adaptive Moment Estimation), который адаптивно настраивает скорость обучения для каждого параметра на основе оценок первого и второго моментов градиента.

Функция потерь categorical_crossentropy измеряет расхождение между истинными метками классов и предсказанными моделью вероятностями. Минимизация данной функции потерь позволяет модели обучаться таким образом, чтобы предсказанные вероятности максимально соответствовали истинным меткам классов. Она определяется следующим образом:

$$L(y, \hat{y}) = - \sum_i y_i \log(\hat{y}_i) \quad (2.3)$$

где y – истинные метки классов;

\hat{y} – предсказанные моделью вероятности,

i – индекс класса.

В процессе обучения модель итеративно проходит через множество обучающих примеров, состоящих из признаков и соответствующих им меток эмоций. На каждой итерации модель делает предсказания, вычисляет ошибку с помощью функции потерь categorical_crossentropy и обновляет свои параметры с помощью оптимизатора Adam, стремясь уменьшить эту ошибку. Градиенты вычисляются с помощью алгоритма обратного распространения ошибки (backpropagation), который распространяет ошибку от выходного слоя к входному, корректируя веса и смещения на каждом слое.

Постепенно, через множество итераций обучения, модель настраивает свои параметры таким образом, чтобы минимизировать ошибку на обучающих

данных. Она учится извлекать информативные признаки и классифицировать эмоции на основе этих признаков.

После завершения обучения модель может быть использована для распознавания эмоций в речевых сигналах, не входивших в обучающую выборку. При подаче нового речевого сигнала на вход обученной модели она проходит через все слои, извлекает признаки и на основе полученных признаков выдает вектор вероятностей принадлежности к каждой эмоции. Эмоция с наибольшей вероятностью считается распознанной моделью для данного сигнала.

Таким образом, предложенная архитектура сверточной нейронной сети позволяет решать задачу распознавания эмоций в человеческой речи. Использование функций активации ReLU и softmax вносит нелинейность в процесс обработки данных и позволяет модели обучаться сложным и абстрактным признакам.

2.4 Оценка качества модели

Для оценки качества модели классификации эмоций использовались стандартные метрики, применяемые для задач многоклассовой классификации: accuracy, recall, precision и F1-score.

В первой части исследования модель будет обучена на трех разных моделях искусственной нейронной сети на объединенных фиксированных данных из четырех датасетов: CREMA-D, RAVDESS, SAVEE и TESS. Для обеспечения корректности оценки, из каждого набора данных будет предварительно выделена фиксированная тестовая выборка в размере 10 записей для каждой эмоции каждого набора данных, которая не будет использоваться в процессе обучения. Это покажет, насколько хорошо модель справится с классификацией эмоций на данных той же структуры и с похожими акустическими характеристиками.

Далее, для более глубокого анализа, будет проведена перекрестная проверка на тех же наборах данных. Будет обучена модель на комбинации из трех наборов данных, а оценка ее качества на четвертом. Этот процесс повторяется для всех возможных комбинаций наборов данных и для одной модели. Такой подход позволит оценить, насколько хорошо модель обобщает знания и насколько точно она может распознавать эмоции на данных, которые не были использованы при обучении.

Анализ полученных результатов будет проводиться в разрезе поставленных задач исследования. Будет проведено сравнение эффективности обучения модели на различных выборках данных, а также будет проведен анализ зависимости эффективности модели от архитектуры CNN и характеристик обучающих выборок. На основе проведенного анализа будут сформулированы выводы о степени влияния исследуемых факторов на качество обучения модели,

а также предложены возможные пути дальнейшего совершенствования методов формирования обучающих выборок и архитектуры нейросетевой модели для повышения точности распознавания эмоций в речи.

2.5 Описание вычислительной платформы и инструментария

Для проведения экспериментов и обучения модели использовался персональный компьютер с операционной системой Windows 11, оснащенный видеокартой NVIDIA GeForce RTX 2060 Super. Эта видеокарта поддерживает технологии CUDA (Compute Unified Device Architecture) и cuDNN (CUDA Deep Neural Network library), разработанные компанией NVIDIA. CUDA представляет собой программно-аппаратную архитектуру параллельных вычислений, которая позволяет использовать графический процессор (GPU) для выполнения ресурсоемких задач — обучения моделей искусственной нейронной сети. cuDNN — это библиотека примитивов глубокого обучения, оптимизированная для GPU, которая содержит высокоэффективные реализации основных операций, используемых в нейронных сетях. Применение этих технологий позволило значительно ускорить процесс обучения модели распознавания эмоций за счет распараллеливания вычислений на графическом процессоре.

В качестве основного языка программирования был выбран Python 3.10, так как он имеет множество удобных библиотек для работы с данными, машинным обучением и нейронными сетями. Разработка кода велась в среде разработки Visual Studio Code, которая обладает удобным интерфейсом и поддержкой множества плагинов.

Обработка и анализ данных производились с помощью библиотек NumPy и Pandas. NumPy использовалась для работы с большими массивами данных и выполнения числовых операций, а Pandas — для создания и управления структурами данных в виде таблиц DataFrame. Их использование позволяет эффективно вычислять и преобразовывать данные, а также организовывать данные в удобной форме с целью их анализа.

Для работы с аудиоданными и извлечения признаков использовалась библиотека librosa. Она позволяет загружать аудиофайлы, вычислять различные признаки и выполнять другие операции.

При разработке нейросетевой модели использовался фреймворк TensorFlow и его высокоуровневый API Keras. Они предоставляют удобные инструменты для создания и обучения нейронных сетей, а также содержат множество готовых компонентов.

Для построения графиков точности и потерь во время обучения модели применялся Matplotlib, а seaborn — для визуализации матрицы ошибок (confusion matrix), которая позволяет наглядно оценить качество классификации модели для каждого класса эмоций. Подробный отчет о процессе обучения и

тестирования модели был сохранен в PDF-файл с помощью класса PdfPages из библиотеки matplotlib.

Библиотека scikit-learn использовалась для разделения данных на обучающие и валидационные наборы, масштабирования данных и вычисления метрик качества модели, таких как accuracy_score и classification_report. Они позволяют вычислять различные метрики для каждого класса эмоций, а также получить общую оценку эффективности модели.

Это были основные библиотеки, которые использовались в скриптах. Теперь кратко рассмотрим другие:

- argparse — работа с аргументами командной строки в скриптах;
- glob — поиск файлов и директорий с использованием шаблонов;
- joblib — параллельная обработка данных для извлечения признаков;
- shutil — работа с файловыми системами, копирование данных для разделения на выборки;
- os — взаимодействие с операционной системой, включая создание и управление директориями, а также проверку существования файлов и директорий;
- re — работа с регулярными выражениями, создание шаблонов для отнесения аудиофайла к определенному набору данных, а также шаблонов для обработки аргументов командной строки;
- pickle — сохранение промежуточных результатов, таких как масштабировщики данных и кодировщики меток, а также их загрузка в другом скрипте;
- random — перемешивание данных и выбор случайных подмножеств для создания тренировочных и тестовых данных;
- io — захват и перенаправление вывода для сохранения информации о модели в отчет.

Таким образом, используя современное оборудование и мощные библиотеки для анализа данных и машинного обучения, была создана эффективная среда для разработки и исследования модели распознавания эмоций в речи.

2.6 Описание программного инструментария

В рамках данной работы разработан программный инструментарий для решения поставленной задачи. Он реализован в виде набора скриптов на языке программирования Python, каждый из которых выполняет определенную функцию в процессе обработки данных, обучения модели и ее оценки, обеспечивая полный цикл работы с данными, начиная от их предварительной подготовки, заканчивая формированием отчетов о проделанной работе. На рис. 2.2 представлен порядок выполнения скриптов.

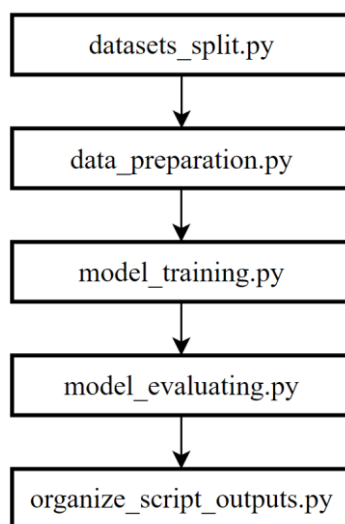


Рисунок 2.2 Порядок выполнения скриптов

Процесс начинается с разделения исходных данных на тренировочные и тестовые подмножества, реализованного в скрипте *datasets_split.py*. Скрипт загружает выбранные аудиофайлы, анализирует имена файлов для определения эмоции, распределяет файлы по двум подмножествам, одно из которых содержит заданное количество файлов для каждой эмоции каждого набора данных, а другое — оставшиеся файлы. Разделение делается для того, чтобы выделить тестовую выборку, которая не будет участвовать в процессе обучения модели, но будет участвовать в оценке. Аргументы `--datasets`, `--count_per_class` и `--output_folder` используются для указания путей к наборам данных, количества файлов каждой эмоции для каждого набора данных в тестовой выборке и пути для сохранения результатов соответственно. В результате выполнения скрипта получаем две папки, представленные на рис. 2.3, первая из которых содержит тренировочное подмножество (обучающую и валидационную выборки), а вторая — тестовую выборку.



Имя	Дата изменения	Тип
 subset1	03.06.2024 15:23	Папка с файлами
 subset2	03.06.2024 15:23	Папка с файлами

Рисунок 2.3 Пример разделенных данных на подмножества

Следующий шаг — подготовка данных для обучения модели, выполняемая скриптом *data_preparation.py*. Скрипт загружает выбранные аудиофайлы, извлекает признаки (MFCC, ZCR, RMSE), создает DataFrame с путями к файлам и метками эмоций, производит масштабирование данных (StandardScaler) и кодирование эмоций (OneHotEncoder), а также создает отчет о подготовке данных. Аргументы `--datasets`, `--output_folder` и `--emotions` используются для указания путей к наборам данных, пути для сохранения результатов и списка распознаваемых эмоций, который можно менять в соответствии с распознаванием желаемых эмоций. В результате получают следующие файлы:

scaler.pickle (файл масштабирования данных), *encoder.pickle* (файл кодирования меток эмоций), *dataset_splits.npz* (файл с разделением данных на обучающие и валидационные наборы), *features.csv* (файл с извлеченными признаками для каждого аудиофайла), *data_preparation.txt* (отчет о подготовке данных).

Скрипт *model_training.py* отвечает за обучение модели. В нем загружаются подготовленные данные, задается модель нейронной сети. Процесс обучения использует методы обратного вызова для контроля точности, ранней остановки и корректировки скорости обучения. Обученная модель сохраняется в файл и формируется отчет, включающий графики точности и потерь, матрицу ошибок и отчет о классификации. Аргументы `--data_file`, `--emotions` и `--output_folder` используются для указания пути к файлу с данными из этапа подготовки данных, списка эмоций и пути для сохранения результатов. В результате получаются следующие файлы: *cnn_model.h5* (файл обученной модели), *model_training_report.pdf* (отчет об обучении модели), *train_paths_with_emotions.txt* (файл с путями к обучающим данным и их метками), *test_paths_with_emotions.txt* (файл с путями к тестовым данным и их метками), *test_predictions_report.txt* (отчет с предсказанными эмоциями для тестовой выборки).

Оценка производительности обученной модели производится скриптом *model_evaluating.py*. В нем загружается модель, файлы с масштабированием и кодированием меток, а затем модель предсказывает эмоции для каждого файла. Сравнивая предсказанные эмоции с истинными метками, скрипт формирует отчет с точностью модели, отчетом с классификацией и детальными результатами для каждого файла (распределениями вероятностей по эмоциям). Аргументы `--data_folder`, `--model_path`, `--pickle_path`, `--emotions` и `--output_folder` используются для указания пути к текстовым файлам, файлам масштабирования и кодирования, списка эмоций и пути для сохранения результатов. В результате получаются файл *model_evaluating_report.txt* (отчет с результатами оценки производительности модели).

Последний этап — организация результатов работы всех скриптов, выполняемая скриптом *organize_script_outputs.py*. Он анализирует файл *run_scripts.bat*, который содержит строки с запуском каждого скрипта, и извлекает информацию об их аргументах, формируя структуру папки для сохранения всех результатов в формате «наборы_данных_для_обучения__наборы_данных_для_оценки__эмоции». Также создается файл *script_run_info.txt* с информацией о запуске. Аргументы `--input_folder` и `--output_folder` используются для указания пути к папке с временными файлами полного цикла работы программы и пути для сохранения организованных результатов. Также в результате создается файл *script_run_info.txt* с запуском всех скриптов и их аргументов. Пример содержимого папки со всеми результатами после выполнения скрипта показан на рис. 2.4, а пример набора папок с сохраненными в них результатами представлен на рис. 2.5.













Имя	Дата изменения	Тип	Размер
 cnn_model.h5	05.06.2024 0:26	Файл "H5"	47 872 КБ
 data_preparation_report.txt	05.06.2024 0:26	Текстовый докум...	1 КБ
 dataset_splits.npz	05.06.2024 0:26	Файл "NPZ"	478 056 КБ
 encoder.pickle	05.06.2024 0:26	Файл "PICKLE"	1 КБ
 features.csv	05.06.2024 0:26	Файл Microsoft Ex...	857 388 КБ
 model_evaluating_report.txt	05.06.2024 0:26	Текстовый докум...	58 КБ
 model_training_report.pdf	05.06.2024 0:26	Microsoft Edge PD...	42 КБ
 scaler.pickle	05.06.2024 0:26	Файл "PICKLE"	39 КБ
 script_run_info.txt	05.06.2024 0:26	Текстовый докум...	1 КБ
 test_paths_with_emotions.txt	05.06.2024 0:26	Текстовый докум...	493 КБ
 test_predictions_report.txt	05.06.2024 0:26	Текстовый докум...	1 551 КБ
 train_paths_with_emotions.txt	05.06.2024 0:26	Текстовый докум...	1 974 КБ

Рисунок 2.4 Пример содержимого папки со всеми результатами










Имя
 CREMA-D__RAVDESS__Anger+Disgust+Fear+Happy+Neutral+Sad
 CREMA-D__SAVEE__Anger+Disgust+Fear+Happy+Neutral+Sad
 CREMA-D__TESS__Anger+Disgust+Fear+Happy+Neutral+Sad
 CREMA-D__CREMA-D__Anger+Disgust+Fear+Happy+Neutral+Sad
 CREMA-D+RAVDESS__SAVEE__Anger+Disgust+Fear+Happy+Neutral+Sad
 CREMA-D+RAVDESS__TESS__Anger+Disgust+Fear+Happy+Neutral+Sad
 CREMA-D+SAVEE__RAVDESS__Anger+Disgust+Fear+Happy+Neutral+Sad
 CREMA-D+RAVDESS+SAVEE__TESS __Anger+Disgust+Fear+Happy+Neutral+Sad1
 RAVDESS+SAVEE+TESS__CREMA-D__Anger+Disgust+Fear+Happy+Neutral+Sad

Рисунок 2.5 Пример набора папок с сохраненными в них результатами

Запуск всех скриптов автоматизирован с помощью файла *run_scripts.bat*, содержащего последовательность команд для выполнения полного цикла обработки данных, обучения и оценки модели. Пример файла с запуском скриптов показан на рис. 2.6.

```
@echo off

python data_preparation.py --datasets datasets\CREMA-D datasets\RAVDESS datasets\TESS --
output_folder "assets" --emotions Anger Disgust Fear Happy Sad
python model_training.py --data_file "assets\dataset_splits.npz" --emotions Anger Disgust Fear
Happy Sad --output_folder "assets"
python model_evaluating.py --data_folder "datasets\SAVEE" --model_path "assets
\cnn_model.h5" --pickle_path "assets" --emotions Anger Disgust Fear Happy Sad --output_folder
"assets"
python organize_script_outputs.py --input_folder "assets" --output_folder "results"

python datasets_split.py --datasets datasets\CREMA-D datasets\RAVDESS datasets\SAVEE datasets
\TESS --count_per_class 10 --output_folder "splited_dataset"
python data_preparation.py --datasets splited_dataset\subset1\CREMA-D splited_dataset\subset1
\RAVDESS splited_dataset\subset1\SAVEE splited_dataset\subset1\TESS --output_folder
"assets" --emotions Anger Disgust Fear Happy Sad
python model_training.py --data_file "assets\dataset_splits.npz" --emotions Anger Disgust Fear
Happy Sad --output_folder "assets"
python model_evaluating.py --data_folder "splited_dataset\subset2" --model_path "assets
\cnn_model.h5" --pickle_path "assets" --emotions Anger Disgust Fear Happy Sad --output_folder
"assets"
python organize_script_outputs.py --input_folder "assets" --output_folder "results"
```

Рисунок 2.6 Пример файла с запуском скриптов

Таким образом, был создан инструментарий для автоматизации всего процесса исследования и проведения комплексного анализа задачи распознавания эмоций в человеческой речи. Гибкая настройка параметров и возможность использования различных наборов данных и моделей делают инструментарий полезным для проведения дальнейших исследований в данной области, а также для разработки прикладных решений.

3 ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

3.1 Обучение и оценка эффективности модели

Для выявления зависимости результата от модели, рассмотрим две модели и выберем одну с лучшим результатом. Сначала все наборы данных были разделены на две выборки: первая, которая в дальнейшем делится на обучающую и валидационную, и вторая часть — тестовая. В данном случае тестовая выборка состоит из аудиозаписей, которые не включены в обучающую и валидационную выборки, при этом они имеют ту же структуру, так как взяты из каждого имеющегося набора данных. Все модели были обучены и протестированы на одних и тех же объединенных аудиофайлах из четырех наборов данных: CREMA-D, RAVDESS, SAVEE и TESS.

В результате подготовки данных было выделено пять сбалансированных классов эмоций: Anger, Disgust, Fear, Happy и Sad, а также были извлечены признаки: MFCC, ZCR и RMSE. Каждый класс содержит по 1883 размеченных аудиофайла. Все данные были случайным образом разделены на обучающую, валидационную (в размере 20% файлов от каждой эмоции каждого набора данных) и тестовую (по 10 файлов каждой эмоции каждого набора данных, итого 200 файлов) выборки.

Первая модель, изображенная на рис. 3.1, состоит из 21 слоя и 4076677 параметров, 4072837 из которых обучаемые. На обучающей выборке на рис. 3.2 модель показала 98.39%. Из рис. 3.3 можно сделать вывод, что модель не склонна к переобучению и демонстрирует хорошую сходимость. Рассмотрим результаты, которые получились хуже всего. Матрица ошибок на рис. 3.4 показала, что модель распознала 19 записей с эмоцией Sad как Fear и 11 наоборот, Anger как Happy 11 записей и 9 наоборот, а также 12 записей Fear как Happy.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 1620, 512)	3072
batch_normalization (Batch Normalization)	(None, 1620, 512)	2048
max_pooling1d (MaxPooling1D)	(None, 810, 512)	0
conv1d_1 (Conv1D)	(None, 810, 512)	1311232
batch_normalization_1 (Batch Normalization)	(None, 810, 512)	2048
max_pooling1d_1 (MaxPooling1D)	(None, 405, 512)	0
dropout (Dropout)	(None, 405, 512)	0
conv1d_2 (Conv1D)	(None, 405, 256)	655616
batch_normalization_2 (Batch Normalization)	(None, 405, 256)	1024
max_pooling1d_2 (MaxPooling1D)	(None, 203, 256)	0
conv1d_3 (Conv1D)	(None, 203, 256)	327936
batch_normalization_3 (Batch Normalization)	(None, 203, 256)	1024
max_pooling1d_3 (MaxPooling1D)	(None, 102, 256)	0
dropout_1 (Dropout)	(None, 102, 256)	0
conv1d_4 (Conv1D)	(None, 102, 128)	98432
batch_normalization_4 (Batch Normalization)	(None, 102, 128)	512
max_pooling1d_4 (MaxPooling1D)	(None, 51, 128)	0
dropout_2 (Dropout)	(None, 51, 128)	0
flatten (Flatten)	(None, 6528)	0
dense (Dense)	(None, 256)	1671424
batch_normalization_5 (Batch Normalization)	(None, 256)	1024
dense_1 (Dense)	(None, 5)	1285
Total params: 4,076,677		
Trainable params: 4,072,837		
Non-trainable params: 3,840		

Рисунок 3.1 Первая модель искусственной нейронной сети

	precision	recall	f1-score	support
Anger	0.9854	0.9841	0.9847	1505
Disgust	0.9900	0.9894	0.9897	1505
Fear	0.9762	0.9801	0.9781	1505
Happy	0.9834	0.9821	0.9827	1505
Sad	0.9847	0.9841	0.9844	1505
accuracy			0.9839	7525
macro avg	0.9839	0.9839	0.9839	7525
weighted avg	0.9839	0.9839	0.9839	7525

Рисунок 3.2 Отчет классификации на обучающей выборке

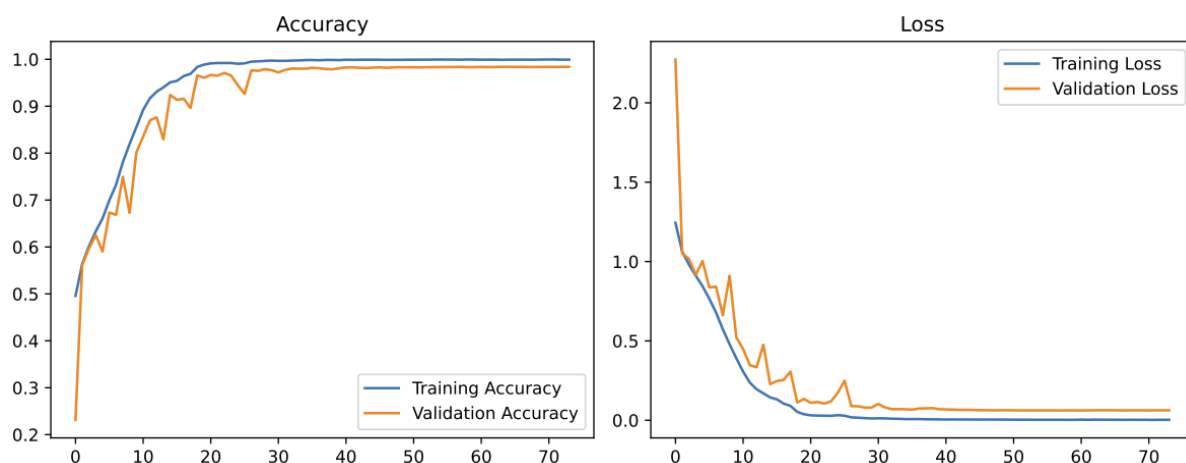


Рисунок 3.3 Графики точности и потерь на обучающей и валидационной выборках

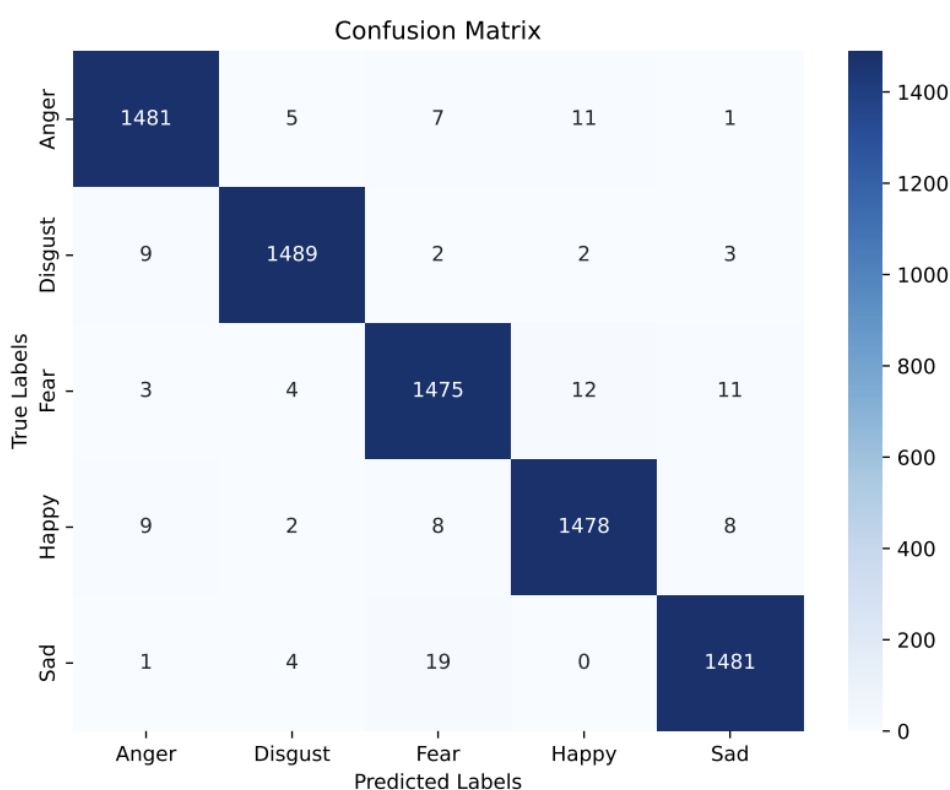


Рисунок 3.4 Матрица ошибок на обучающей выборке

Рис. 3.5 демонстрирует, что на тестовой выборке количество правильно предсказанных эмоций составило 130 из 200, что составляет 65% точности. Точность и полнота варьируются в зависимости от класса, что указывает на неравномерное распределение различных эмоций. Например, для класса Anger точность составляет 58.82%, полнота 75%, F1-мера 65.93%, что говорит о том, что модель лучше распознает случаи злости, чем другие эмоции. У класса Disgust эти показатели достаточно равномерные, соответственно, хорошо распознается. С классами Fear и Happy у модели возникают трудности в распознавании, что следует из самых низких показателей F1-меры в 60.87% и 60% соответственно.

Number of correctly predicted emotions: 130 out of 200
Overall accuracy: 65.0000%

Detailed Classification Report for the test set:

	precision	recall	f1-score	support
Anger	0.5882	0.7500	0.6593	40
Disgust	0.7568	0.7000	0.7273	40
Fear	0.7241	0.5250	0.6087	40
Happy	0.5400	0.6750	0.6000	40
Sad	0.7273	0.6000	0.6575	40
accuracy			0.6500	200
macro avg	0.6673	0.6500	0.6506	200
weighted avg	0.6673	0.6500	0.6506	200

Рисунок 3.5 Отчет классификации на тестовой выборке

Предположим, что модель плохо классифицирует эмоции. Улучшим модель, добавив вместо 256 нейронов 512 в полносвязный слой, а также добавим новые слои и проверим, улучшится ли результат.

Вторая модель на рис. 3.6, по сравнению с прошлой моделью, состоит из большего числа слоев в размере 29, а также большего числа параметров — 13236357, 13227909 из которых обучаемые. Точность на обучающей выборке на рис. 3.7 составляет 98.86%, что больше, чем у прошлых результатов. Проблемы у модели с эмоциями Fear и Anger. Из рис. 3.8 можно сделать вывод, что данная модель также не склонна к переобучению и демонстрирует хорошую сходимость. Из матрицы ошибок на рис. 3.9 следует, что модель спутала Fear и Disgust 8 раз, а Sad и Disgust 9 раз, наоборот — 8 раз. Класс Fear был распознан как Happy 9 раз, а Anger как Fear — 8 раз.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 1620, 1024)	8192
batch_normalization (Batch Normalization)	(None, 1620, 1024)	4096
max_pooling1d (MaxPooling1D)	(None, 810, 1024)	0
conv1d_1 (Conv1D)	(None, 810, 1024)	7341056
batch_normalization_1 (Batch Normalization)	(None, 810, 1024)	4096
max_pooling1d_1 (MaxPooling1D)	(None, 405, 1024)	0
dropout (Dropout)	(None, 405, 1024)	0
conv1d_2 (Conv1D)	(None, 405, 512)	2621952
batch_normalization_2 (Batch Normalization)	(None, 405, 512)	2048
max_pooling1d_2 (MaxPooling1D)	(None, 203, 512)	0
conv1d_3 (Conv1D)	(None, 203, 512)	1311232
batch_normalization_3 (Batch Normalization)	(None, 203, 512)	2048
max_pooling1d_3 (MaxPooling1D)	(None, 102, 512)	0
dropout_1 (Dropout)	(None, 102, 512)	0
conv1d_4 (Conv1D)	(None, 102, 256)	655616
batch_normalization_4 (Batch Normalization)	(None, 102, 256)	1024
max_pooling1d_4 (MaxPooling1D)	(None, 51, 256)	0
conv1d_5 (Conv1D)	(None, 51, 256)	327936
batch_normalization_5 (Batch Normalization)	(None, 51, 256)	1024
max_pooling1d_5 (MaxPooling1D)	(None, 26, 256)	0
dropout_2 (Dropout)	(None, 26, 256)	0
conv1d_6 (Conv1D)	(None, 26, 128)	98432
batch_normalization_6 (Batch Normalization)	(None, 26, 128)	512
max_pooling1d_6 (MaxPooling1D)	(None, 13, 128)	0
dropout_3 (Dropout)	(None, 13, 128)	0
flatten (Flatten)	(None, 1664)	0
dense (Dense)	(None, 512)	852480
batch_normalization_7 (Batch Normalization)	(None, 512)	2048
dense_1 (Dense)	(None, 5)	2565
Total params: 13,236,357		
Trainable params: 13,227,909		
Non-trainable params: 8,448		

Рисунок 3.6 Третья модель искусственной нейронной сети

	precision	recall	f1-score	support
Anger	0.9960	0.9854	0.9906	1505
Disgust	0.9829	0.9927	0.9878	1505
Fear	0.9926	0.9827	0.9876	1505
Happy	0.9881	0.9894	0.9887	1505
Sad	0.9835	0.9927	0.9881	1505
accuracy			0.9886	7525
macro avg	0.9886	0.9886	0.9886	7525
weighted avg	0.9886	0.9886	0.9886	7525

Рисунок 3.7 Отчет классификации третьей модели на обучающей выборке

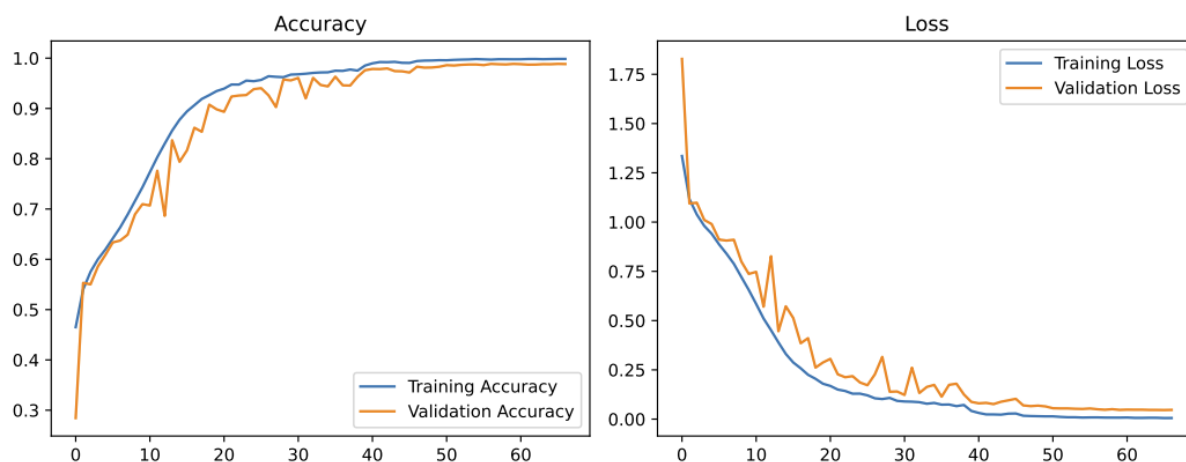


Рисунок 3.8 Графики точности и потерь на обучающей и валидационной выборках третьей модели

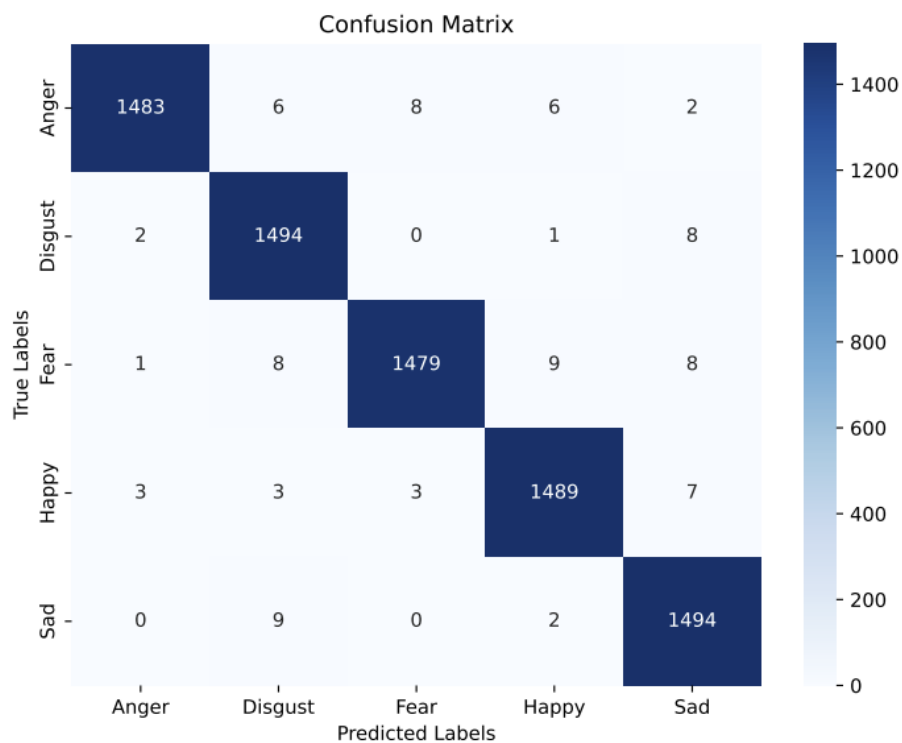


Рисунок 3.9 Матрица ошибок на обучающей выборке

В итоге получаем, из рис. 3.10, что количество правильно предсказанных эмоций составило 130 из 200, общая точность составила 65%, как и в первой модели. Лучший результат распознавания у Anger, где точность 68.09%, полнота 80% и F1-мера 73.56%, что говорит о путанице других классов с Anger. Также хороший результат у Fear, показавший точность 64.1%, полноту 62.5% и F1-меру 62.29%, где показатели достаточно хорошо распределены. Классы Disgust и Sad имеют самые низкие показатели F1-меры — 61.18% и 61.11% соответственно, а из анализа метрик точности и полноты следует, что модель чаще путает Disgust с другими классами, а другие классы воспринимает как Sad.

Number of correctly predicted emotions: 130 out of 200				
Overall accuracy: 65.0000%				
Detailed Classification Report for the test set:				
	precision	recall	f1-score	support
Anger	0.6809	0.8000	0.7356	40
Disgust	0.5778	0.6500	0.6118	40
Fear	0.6410	0.6250	0.6329	40
Happy	0.6757	0.6250	0.6494	40
Sad	0.6875	0.5500	0.6111	40
accuracy			0.6500	200
macro avg	0.6526	0.6500	0.6482	200
weighted avg	0.6526	0.6500	0.6482	200

Рисунок 3.10 Отчет классификации третьей модели на тестовой выборке

Из полученных результатов можно сделать вывод, что изменение структуры и внутренних параметров сверточной нейронной сети не повлияло на результат при текущих выборках. Тогда для дальнейшего тестирования выберем первую модель. Также полученный результат может быть из-за данных, и для того, чтобы это проверить, необходимо протестировать выбранную модель на различных наборах данных, которые полностью не используются при обучении, но используются при тестировании.

3.2 Тестирование модели на различных наборах данных

Теперь важно посмотреть, какие результаты даст модель, которая будет тестироваться на данных из совсем другого набора данных. То есть модель будет обучаться на комбинации из трех наборов данных, а затем тестироваться на оставшемся четвертом, который не использовался при обучении. Это позволяет сделать выводы о влиянии наборов данных на обучение.

Оценка будет производиться при помощи одной модели, представленной на рис. 3.11, которая была рассмотрена и выбрала в прошлом разделе.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 1620, 512)	3072
batch_normalization (Batch Normalization)	(None, 1620, 512)	2048
max_pooling1d (MaxPooling1D)	(None, 810, 512)	0
conv1d_1 (Conv1D)	(None, 810, 512)	1311232
batch_normalization_1 (Batch Normalization)	(None, 810, 512)	2048
max_pooling1d_1 (MaxPooling1D)	(None, 405, 512)	0
dropout (Dropout)	(None, 405, 512)	0
conv1d_2 (Conv1D)	(None, 405, 256)	655616
batch_normalization_2 (Batch Normalization)	(None, 405, 256)	1024
max_pooling1d_2 (MaxPooling1D)	(None, 203, 256)	0
conv1d_3 (Conv1D)	(None, 203, 256)	327936
batch_normalization_3 (Batch Normalization)	(None, 203, 256)	1024
max_pooling1d_3 (MaxPooling1D)	(None, 102, 256)	0
dropout_1 (Dropout)	(None, 102, 256)	0
conv1d_4 (Conv1D)	(None, 102, 128)	98432
batch_normalization_4 (Batch Normalization)	(None, 102, 128)	512
max_pooling1d_4 (MaxPooling1D)	(None, 51, 128)	0
dropout_2 (Dropout)	(None, 51, 128)	0
flatten (Flatten)	(None, 6528)	0
dense (Dense)	(None, 256)	1671424
batch_normalization_5 (Batch Normalization)	(None, 256)	1024
dense_1 (Dense)	(None, 5)	1285
=====		
Total params: 4,076,677		
Trainable params: 4,072,837		
Non-trainable params: 3,840		

Рисунок 3.11 Модель искусственной нейронной сети

Сначала проведем оценку распознавания CREMA-D на модели, которая обучена на наборах данных RAVDESS, SAVEE и TESS. Точность на обучающей выборке на рис. 3.12 составляет 98.85%. Основные сложности возникают при распознавании классов Anger и Happy, так как у них самый низкий показатель F1-меры — 97.67% и 97.25% соответственно. У класса Anger точность 99.01%, полнота 96.35%, а у Happy точность 96.24%, полнота 98.27%, у других классов точность и полнота немного отличаются. Это говорит о том, что Anger и Happy

чаще всего путаются между собой при распознавании. Из матрицы ошибок на рис. 3.14 видно, что модель больше всего ошибается, делая предсказание Anger как Happy в количестве 15 раз, в то время как наоборот всего 1 раз.

	precision	recall	f1-score	support
Anger	0.9901	0.9635	0.9767	521
Disgust	0.9884	0.9827	0.9856	521
Fear	0.9735	0.9885	0.9810	521
Happy	0.9624	0.9827	0.9725	521
Sad	0.9904	0.9866	0.9885	521
accuracy			0.9808	2605
macro avg	0.9810	0.9808	0.9808	2605
weighted avg	0.9810	0.9808	0.9808	2605

Рисунок 3.12 Отчет классификации модели на обучающей выборке

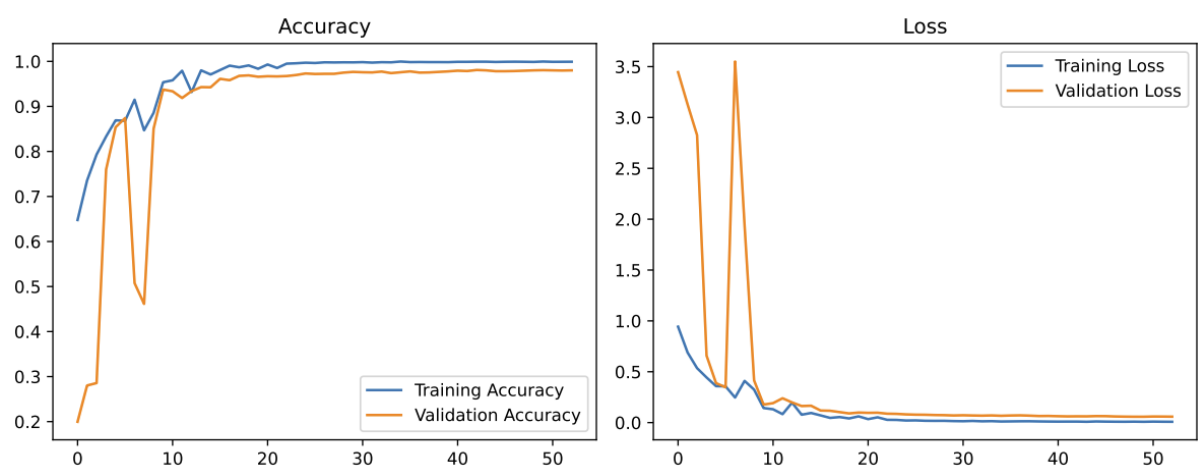


Рисунок 3.13 Графики точности и потерь на обучающей и валидационной выборках модели

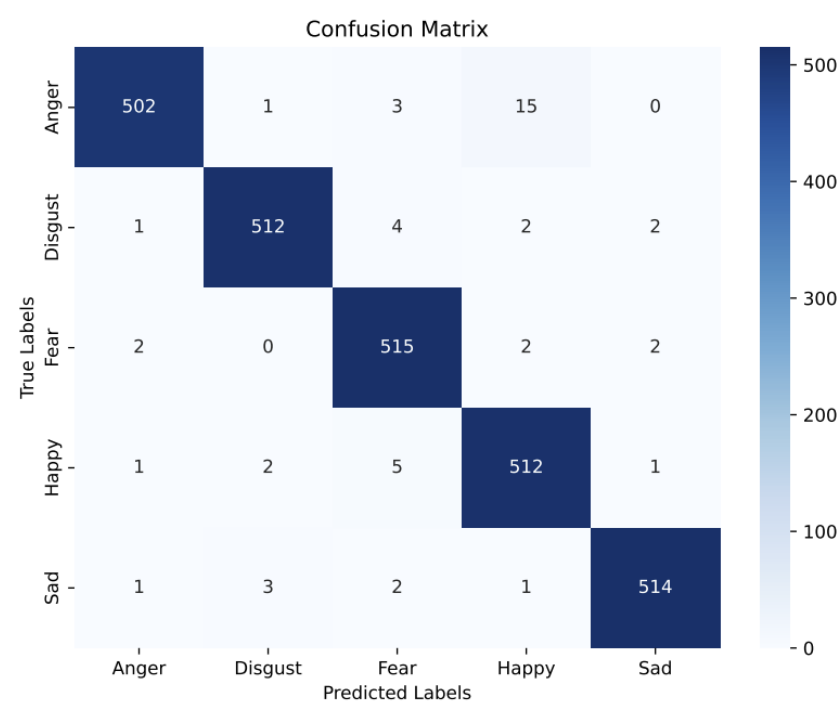


Рисунок 3.14 Матрица ошибок на обучающей выборке

На основе рис. 3.15 можно сделать вывод, что количество правильно предсказанных эмоций составляет 1934 из 6354, что соответствует общей точности в 30.4375%. Наилучший результат распознавания наблюдается у класса Sad, где точность составляет 31.29%, полнота 88.98%, а F1-мера 46.3%. Это свидетельствует о путанице других классов с Sad. Также хороший результат показал класс Anger с точностью 51.45%, полнотой 22.34% и F1-мерой 31.16%. Наихудшие результаты у классов Fear и Happy, где F1-мера составляет 12% и 17.92% соответственно.

Number of correctly predicted emotions: 1934 out of 6354				
Overall accuracy: 30.4375%				
Detailed Classification Report for the test set:				
	precision	recall	f1-score	support
Anger	0.5145	0.2234	0.3116	1271
Disgust	0.2699	0.1786	0.2150	1271
Fear	0.1846	0.0889	0.1200	1271
Happy	0.2439	0.1416	0.1792	1271
Sad	0.3129	0.8898	0.4630	1270
accuracy			0.3044	6354
macro avg	0.3052	0.3045	0.2578	6354
weighted avg	0.3052	0.3044	0.2577	6354

Рисунок 3.15 Отчет классификации модели на тестовой выборке

Теперь проведем оценку распознавания RAVDESS на модели, которая обучена на наборах данных CREMA-D, SAVEE и TESS. Рис. 3.16 демонстрирует, что на обучающей выборке модель достигла общей точности в 98.95%. При анализе результатов классификации получаем, что у каждого класса высокие и равномерные показатели не только F1-меры, но и достаточно равномерные показатели точности и полноты. Матрица ошибок на рис. 3.18 показывает, что модель больше всего ошибается, делая предсказание Happy как Fear в количестве 10 раз, наоборот — 6 раз. Также модель чаще всего путает Fear и Sad — 9 раз, а также Disgust как Sad — 7 раз, и наоборот — 6 раз.

	precision	recall	f1-score	support
Anger	0.9942	0.9899	0.9920	1384
Disgust	0.9899	0.9870	0.9884	1384
Fear	0.9870	0.9870	0.9870	1384
Happy	0.9877	0.9892	0.9884	1384
Sad	0.9885	0.9942	0.9914	1384
accuracy			0.9895	6920
macro avg	0.9895	0.9895	0.9895	6920
weighted avg	0.9895	0.9895	0.9895	6920

Рисунок 3.16 Отчет классификации модели на обучающей выборке

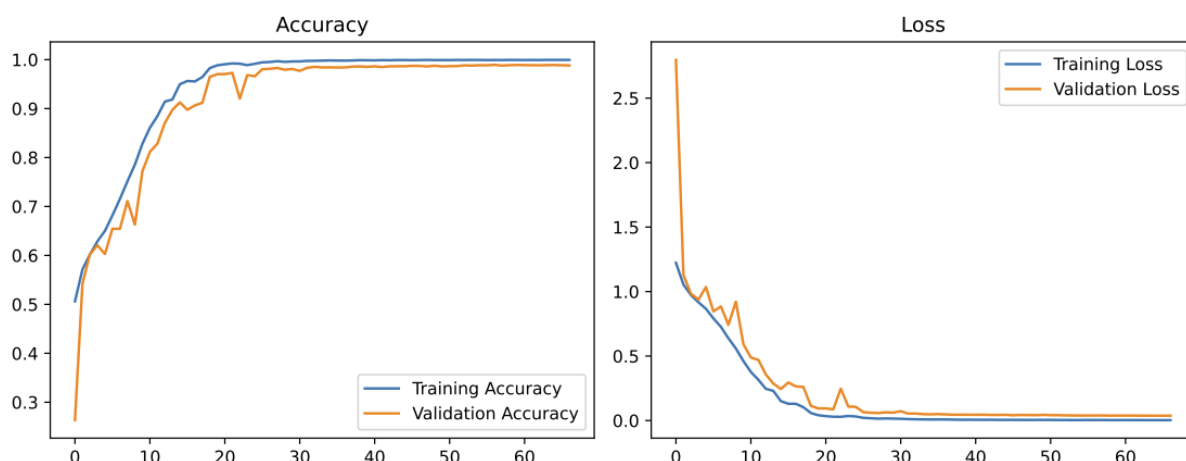


Рисунок 3.17 Графики точности и потерь на обучающей и валидационной выборках модели

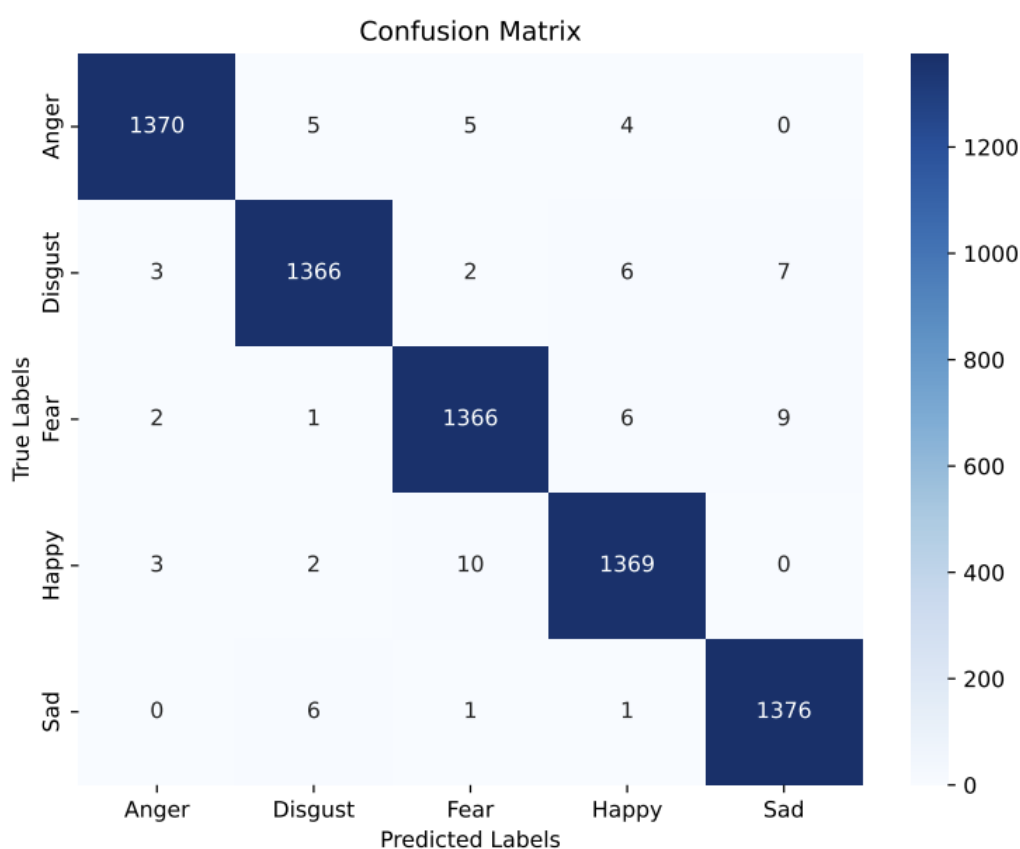


Рисунок 3.18 Матрица ошибок на обучающей выборке

На основе рис. 3.19, количество правильно предсказанных эмоций составляет 295 из 960 записей с общей точностью 30.73%. Самый лучший результат получился у классов Anger и Fear — F1-мера составляет 39.04% и 38.54%. Самыми интересными классами оказались Fear, Happy и Sad. Модель находит 82.29% примеров Fear и имеет низкую точность в 25.16%, что указывает на большое количество ложноположительных срабатываний. Модель обнаруживает всего 9.38% примеров класса Happy, но имеет точность 36.73%, когда предсказывает класс Happy. Также, модель имеет самую высокую точность для класса Sad — 50%, но она находит всего 2.6% примеров Sad.

```

Number of correctly predicted emotions: 295 out of 960
Overall accuracy: 30.7292%

Detailed Classification Report for the test set:

```

	precision	recall	f1-score	support
Anger	0.4610	0.3385	0.3904	192
Disgust	0.3712	0.2552	0.3025	192
Fear	0.2516	0.8229	0.3854	192
Happy	0.3673	0.0938	0.1494	192
Sad	0.5000	0.0260	0.0495	192
accuracy			0.3073	960
macro avg	0.3902	0.3073	0.2554	960
weighted avg	0.3902	0.3073	0.2554	960

Рисунок 3.19 Отчет классификации модели на тестовой выборке

Теперь проведем оценку распознавания SAVEE на модели, которая обучена на наборах данных CREMA-D, RAVDESS и TESS. Из рис. 3.20 получаем, что на обучающей выборке модель достигла общей точности в 98.62%. Анализируя отчет классификации модели, получаем, что у каждого класса высокие и равномерные показатели F1-меры, а также и достаточно равномерные показатели точности и полноты. Из матрицы ошибок на рис. 3.22 следует, что модель больше всего ошибается, делая предсказание Happy как Anger в количестве 17 раз, а также Disgust как Sad в количестве 10. Anger как Fear — 8 ошибочных предсказаний, Happy как Fear и Sad как Fear — 8 и 9 ошибочных предсказаний соответственно.

	precision	recall	f1-score	support
Anger	0.9820	0.9866	0.9843	1489
Disgust	0.9912	0.9879	0.9896	1489
Fear	0.9819	0.9846	0.9832	1489
Happy	0.9885	0.9799	0.9841	1489
Sad	0.9873	0.9919	0.9896	1489
accuracy			0.9862	7445
macro avg	0.9862	0.9862	0.9862	7445
weighted avg	0.9862	0.9862	0.9862	7445

Рисунок 3.20 Отчет классификации модели на обучающей выборке

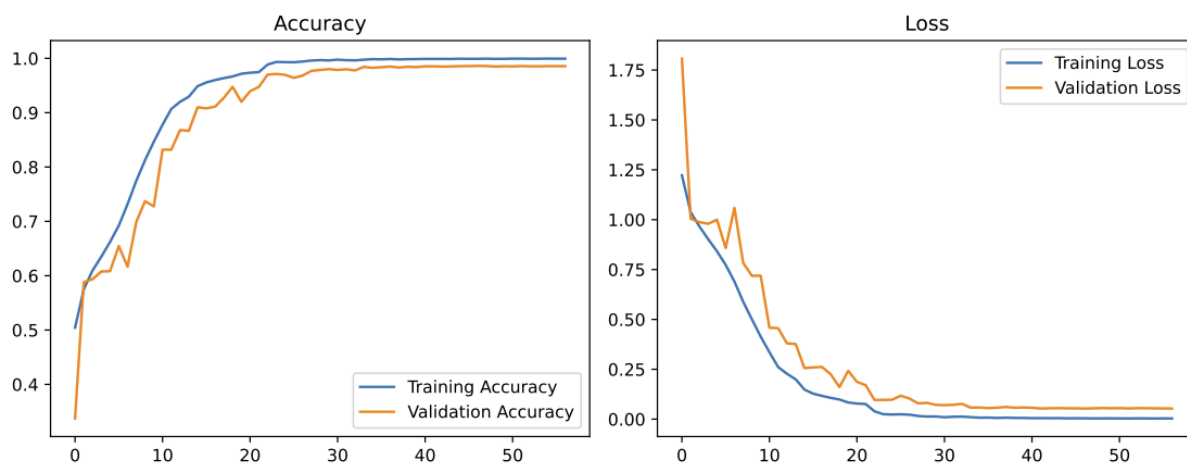


Рисунок 3.21 Графики точности и потерь на обучающей и валидационной выборках модели

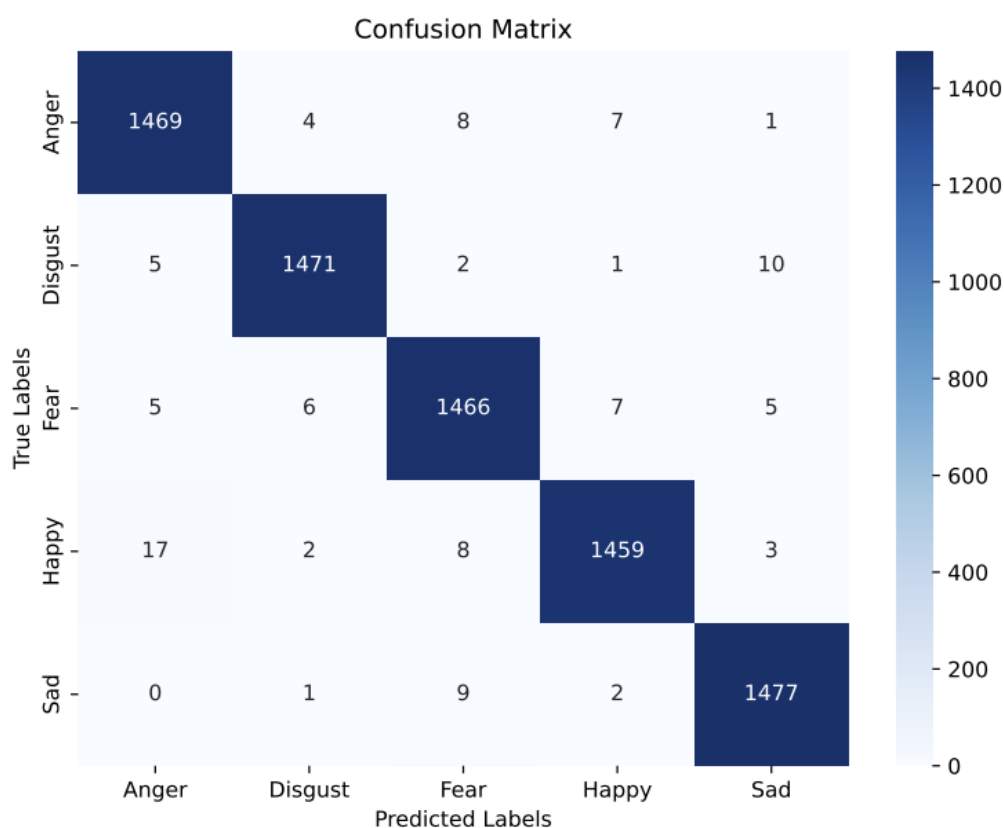


Рисунок 3.22 Матрица ошибок на обучающей выборке

Исходя из рис. 3.23, количество правильно предсказанных эмоций составляет 80 из 300 записей с общей точностью 26.67%. Самый плохой результат оказался у класса Fear — точность 23.81%, полнота 8.33% и F1-мера 12.35%. Классы Anger и Disgust имеют точность ниже, чем полноту — точность 26.85% и 28.09%, полнота 48.33% и 41.67% соответственно. При этом у них лучшая F1-мера — 34.52% и 33.56% соответственно. Классы Happy и Sad имеют точность выше, чем полноту — точность 28.57% и 24.07%, полнота 13.33% и 21.67%, F1-мера 18.18% и 22.81% соответственно.

Number of correctly predicted emotions: 80 out of 300
Overall accuracy: 26.6667%

Detailed Classification Report for the test set:

	precision	recall	f1-score	support
Anger	0.2685	0.4833	0.3452	60
Disgust	0.2809	0.4167	0.3356	60
Fear	0.2381	0.0833	0.1235	60
Happy	0.2857	0.1333	0.1818	60
Sad	0.2407	0.2167	0.2281	60
accuracy			0.2667	300
macro avg	0.2628	0.2667	0.2428	300
weighted avg	0.2628	0.2667	0.2428	300

Рисунок 3.23 Отчет классификации модели на тестовой выборке

Теперь проведем оценку распознавания TESS на модели, которая обучена на наборах данных CREMA-D, RAVDESS и SAVEE. Рис. 3.24 демонстрирует, что на обучающей выборке модель достигла общей точности в 97.81%. Это самая низкая общая точность на обучающей выборке среди всех комбинаций наборов данных для оценки. Матрица ошибок на рис. 3.26 показывает, что модель больше всего ошибается, делая предсказание Happy как Anger в количестве 17 раз, наоборот — 14 раз, а также Sad как Disgust — 11 раз, и Fear как Sad — 10 раз, и наоборот — 7 раз.

	precision	recall	f1-score	support
Anger	0.9698	0.9770	0.9734	1217
Disgust	0.9826	0.9770	0.9798	1217
Fear	0.9835	0.9770	0.9802	1217
Happy	0.9754	0.9762	0.9758	1217
Sad	0.9795	0.9836	0.9815	1217
accuracy			0.9781	6085
macro avg	0.9782	0.9781	0.9781	6085
weighted avg	0.9782	0.9781	0.9781	6085

Рисунок 3.24 Отчет классификации модели на обучающей выборке

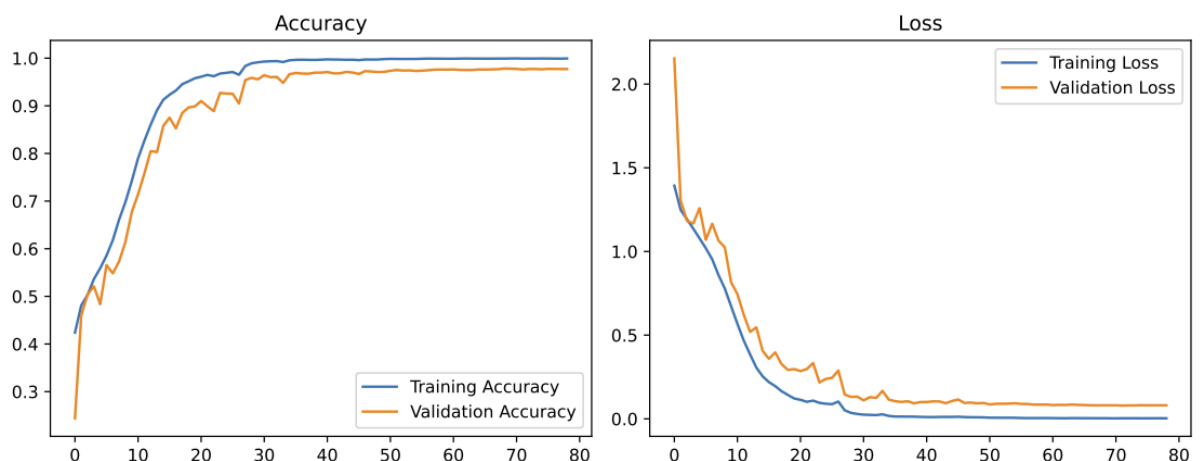


Рисунок 3.25 Графики точности и потерь на обучающей и валидационной выборках модели

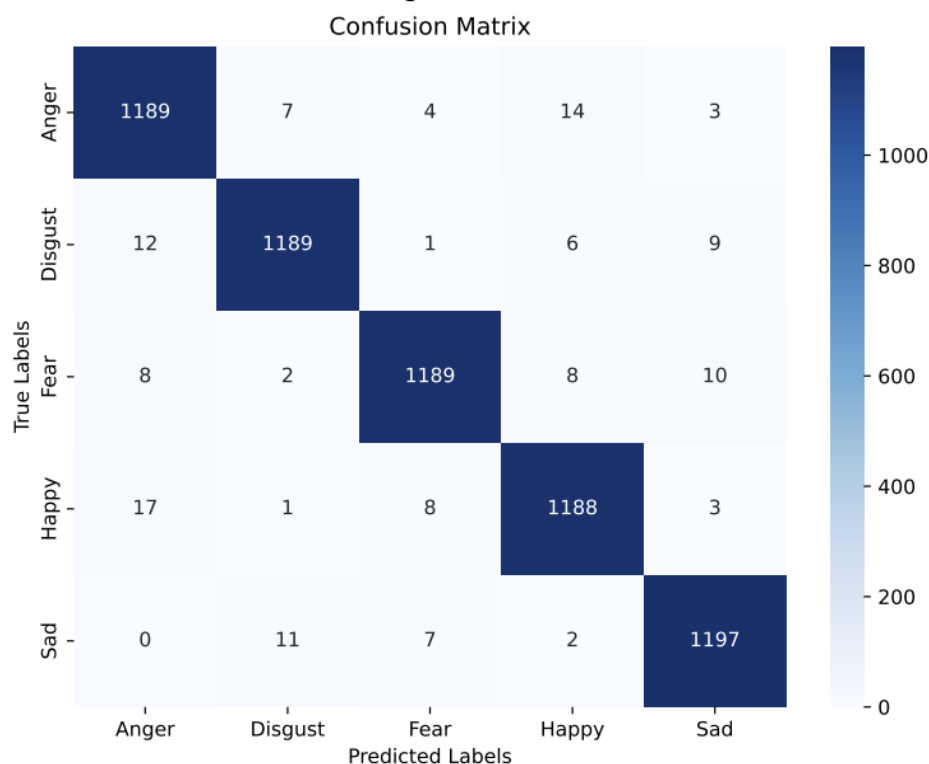


Рисунок 3.26 Матрица ошибок на обучающей выборке

На основе рис. 3.27, количество правильно предсказанных эмоций получается 541 из 2000 записей, общая точность 27.05%. Наилучшие результаты были достигнуты для классов Disgust и Sad, с F1-мерой 29.40% и 30.13% соответственно. Стоит отметить особенности классов Happy и Sad. Модель обнаруживает 39% примеров класса Happy, однако точность предсказания этого класса составляет лишь 24%. С другой стороны, для класса Sad модель демонстрирует самую высокую точность — 36.82%, но находит только 25.5% примеров этого класса.

Number of correctly predicted emotions: 541 out of 2000
Overall accuracy: 27.0500%

Detailed Classification Report for the test set:

	precision	recall	f1-score	support
Anger	0.2334	0.2200	0.2265	400
Disgust	0.2955	0.2925	0.2940	400
Fear	0.2600	0.1950	0.2229	400
Happy	0.2400	0.3900	0.2971	400
Sad	0.3682	0.2550	0.3013	400
accuracy			0.2705	2000
macro avg	0.2794	0.2705	0.2684	2000
weighted avg	0.2794	0.2705	0.2684	2000

Рисунок 3.27 Отчет классификации модели на тестовой выборке

4 АНАЛИЗ РЕЗУЛЬТАТОВ РАБОТЫ

4.1 Оценка качества распознавания эмоций на различных наборах данных

Анализ результатов экспериментов по оценке качества распознавания эмоций на различных наборах данных выявил несколько важных закономерностей и проблем, требующих особого внимания.

Прежде всего, значительное расхождение между высокой точностью модели на валидационных выборках (от 97.81% до 98.95%) и относительно низкой точностью на тестовых наборах (от 26.67% до 30.73%) указывает не только на существенные различия в структуре и характеристиках используемых наборов данных, но и на другие причины.

Высокая точность на валидационных выборках свидетельствует о способности модели эффективно распознавать эмоции в рамках обучающих данных. Однако, значительное снижение точности на тестовых наборах указывает на ее ограниченную способность обобщать знания на новые, ранее не встречавшиеся данные. Это может быть связано с различиями в акустических и лингвистических характеристиках эмоций между наборами данных.

Кроме того, можно заметить неравномерное качество распознавания различных эмоциональных состояний. Классы Sad и Anger показали относительно высокие результаты на большинстве тестовых наборов, в то время как классы Fear и Happy оказались наиболее сложными для распознавания. Это может быть обусловлено несколькими факторами, включая различия в степени выраженности эмоций между наборами данных, возможный дисбаланс классов в обучающих данных и специфические акустические характеристики отдельных эмоций.

Таким образом, результаты экспериментов показывают, что качество распознавания эмоций существенно зависит от используемого набора данных для тестирования. Общая точность варьируется от 26.67% до 30.73%, а F1-мера для отдельных классов эмоций может значительно отличаться в зависимости от набора данных. Это свидетельствует о необходимости учета специфики каждого набора данных при обучении моделей для распознавания эмоций.

4.2 Анализ факторов, влияющих на распознавание эмоций

Распознавание эмоций в речи с помощью глубокого обучения — комплексная задача, эффективность которой зависит от ряда факторов. В данном разделе проведен анализ ключевых аспектов, оказывающих влияние на точность

и надежность разработанных моделей, основываясь на результатах экспериментальной части исследования.

Обучение моделей распознавания эмоций на основе комбинирования наборов данных CREMA-D, RAVDESS, SAVEE, TESS связано с рядом трудностей. Во-первых, наблюдается несогласованность в представлении и интерпретации эмоций, их количестве и нюансах выражения. Во-вторых, значительное расхождение в объеме наборов (от 60 до 1271 записей на эмоцию). В-третьих, различия в структуре высказываний, лексике и стиле речи между наборами затрудняют обобщение акустических признаков эмоций. Разнообразие демографических характеристик дикторов, таких как пол и возраст, в разных наборах может привести к неравномерной чувствительности модели к эмоциям, выраженным разными группами. Дополнительную сложность представляет «театральность» наборов данных с актерами, так как в каждом наборе данных свои преувеличенные представления эмоций, что также делает невозможным применимость модели к анализу естественной речи.

Извлечение информативных признаков из аудиоданных является следующим важным этапом для построения эффективной модели. При этом выбор и настройка параметров извлечения играют решающую роль в качестве получаемых признаков и, как следствие, в точности модели. Различные параметры могут подчеркивать те или иные характеристики аудиосигнала, что делает их более или менее подходящими для выявления определенных эмоций. В рамках данной работы были выбраны параметры извлечения признаков из документации librosa, а также три часто используемых для данной задачи признака. Если брать больше признаков, то модель будет гораздо сложнее. Тем не менее, оптимизация параметров извлечения для конкретной задачи представляет собой перспективное направление дальнейших исследований.

Архитектура нейронной сети также играет ключевую роль. Выбор архитектуры, количества и типа слоев, а также настройка гиперпараметров непосредственно влияют на способность модели улавливать тонкие нюансы эмоциональной окраски речи. Более сложные модели могут показать как лучше результат, так и хуже, что требует поиска оптимального баланса между сложностью и обобщающей способностью.

Таким образом, качество распознавания эмоций в модели зависит от множества факторов, включая характеристики наборов данных, архитектуру модели и методы предобработки. Учет этих факторов и применение соответствующих подходов может значительно улучшить обобщающую способность и устойчивость моделей распознавания эмоций.

ЗАКЛЮЧЕНИЕ

В рамках данной выпускной квалификационной работы была поставлена цель исследовать влияние обучающей выборки и извлекаемых признаков на распознавание эмоций в человеческой речи с применением сверточных нейронных сетей. Для достижения этой цели были выполнены следующие задачи: анализ теоретических основ распознавания эмоций, выбор и подготовка наборов данных, выбор архитектуры сверточной нейронной сети, выбор и извлечение признаков, проведение экспериментов по обучению и тестированию модели, а также анализ полученных результатов.

Основные результаты работы показали, что разработанная модель с использованием наборов данных с актерской речью не продемонстрировала высокой эффективности в задаче распознавания эмоций. Несмотря на достижение высокой точности на обучающей выборке, при тестировании на новых наборах данных, которые не использовались при обучении, её точность существенно снизилась. Это указывает на то, что структура, состав и свойства обучающей выборки оказывают значительное влияние на качество обучения нейросетевой модели.

Качество использованных для обучения модели данных оказалось недостаточным из-за значительных различий в количестве записей, структуре высказываний и стиле речи между наборами данных, а также неравномерного распределения эмоций. Также для улучшения качества распознавания эмоций необходимо обеспечить фонетическую сбалансированность данных, что позволит минимизировать влияние индивидуальных особенностей голоса и произношения на результаты модели.

Также следует отметить важность выбора признаков и настройки их параметров, так как от этого зависит качество распознавания эмоций и точность результата. Различные параметры могут подчеркивать те или иные характеристики аудиосигнала, что делает их более или менее подходящими для выявления определенных эмоций.

В дальнейшем планируется расширить исследование, включив в него следующие направления:

- *вариативность данных*: сбор и подготовка более обширных и разнообразных наборов данных;
- *выбор признаков*: подбор признаков и их параметров для улучшения результатов распознавания;
- *исследование архитектур нейронных сетей*: применение и сравнение различных архитектур для выявления наиболее эффективных подходов к распознаванию эмоций в речи.

Таким образом, данная выпускная квалификационная работа выявила текущие проблемы и определила пути их решения. Полученные результаты служат основой для дальнейших исследований и разработок в области распознавания эмоций в речи.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation / K. Oh, D. Lee, B. Ko, H. Choi // 2017 18th IEEE International Conference on Mobile Data Management (MDM). 2017. P. 371-375. DOI: 10.1109/MDM.2017.64
2. Lee C., Narayanan S. Toward detecting emotions in spoken dialogs // IEEE Transactions on Speech and Audio Processing. 2005. Vol. 13. P. 293-303. DOI: 10.1109/TSA.2004.838534
3. Multimodal Deep Learning for Mental Disorders Prediction from Audio Speech Samples / H. Naderi, B. Soleimani, S. Rempel, S. Matwin, R. Uher [Электронный ресурс] // arxiv.org. 2019. URL: <https://arxiv.org/abs/1909.01067> (дата обращения: 17.03.2024)
4. Speech Emotion Recognition Using Deep Learning Techniques: A Review / R. Khalil, E. Jones, M. Babar, T. Jan, M. Zafar, T. Alhussain // IEEE Access. 2019. Vol. 7. P. 117327-117345. DOI: 10.1109/ACCESS.2019.2936124
5. Sagha H., Deng J., Schuller B. The effect of personality trait, age, and gender on the performance of automatic speech valence recognition // 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). 2017. P. 86-91. DOI: 10.1109/ACII.2017.8273583
6. Articulation constrained learning with application to speech emotion recognition / M. Shah, M. Tu, V. Berisha, C. Chakrabarti, A. Spanias // Eurasip Journal on Audio, Speech, and Music Processing. 2019. DOI: 10.1186/s13636-019-0157-9
7. Wester M., Watts O., Henter G. E. Evaluating comprehension of natural and synthetic conversational speech // Proc. speech prosody. 2016. Vol. 8. P. 736-740
8. Scherer K. R., Banse R., Wallbott H. G. Emotion inferences from vocal expression correlate across languages and cultures // Journal of Cross-cultural psychology. 2001. Vol. 32, № 1. P. 76-92. DOI: 10.1177/0022022101032001009
9. Montenegro C., Maravillas E. Acoustic-prosodic recognition of emotion in speech // 2015 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM). 2015. P. 1-5. DOI: 10.1109/HNICEM.2015.7393229
10. Abdel-Hamid L., Shaker N., Emara I. Analysis of Linguistic and Prosodic Features of Bilingual Arabic–English Speakers for Speech Emotion Recognition // IEEE Access. 2020. Vol. 8. P. 72957-72970. DOI: 10.1109/ACCESS.2020.2987864
11. Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion / B. Al-onazi, M. Nauman, R. Jahangir, M. Malik, E. Alkhamash, A. Elshewey // Applied Sciences. 2022. DOI: 10.3390/app12189188
12. Emotional speech recognition using deep neural networks / L. Trinh Van, T. Dao Thi Le, T. Le Xuan, E. Castelli // Sensors. 2022. Vol. 22, № 4. P. 1414
13. Zhao J., Mao X., Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks // Biomedical signal processing and control. 2019. Vol. 47. P. 312-323. DOI: 10.1016/j.bspc.2018.08.035

14. Bangla Speech-based Emotion Detection using a Hybrid CNN-Transformer Approach / S. H. A. Shuvo, R. Khan // 2023 8th International Conference on Communication, Image and Signal Processing (CCISP). 2023. P. 163-167. DOI: 10.1109/CCISP.2023.8273583
15. Zhang Z., Cummins N., Schuller B. Advanced data exploitation in speech analysis: An overview // IEEE Signal Processing Magazine. 2017. Vol. 34, № 4. P. 107-129. DOI: 10.1109/MSP.2017.2699358
16. Chen J., Ro T., Zhu Z. Emotion recognition with audio, video, EEG, and EMG: a dataset and baseline approaches // IEEE Access. 2022. Vol. 10. P. 13229-13242. DOI: 10.1109/ACCESS.2022.3146729
17. Koduru A., Valiveti H. B., Budati A. K. Feature extraction algorithms to improve the speech emotion recognition rate // International Journal of Speech Technology. 2020. Vol. 23, № 1. P. 45-55. DOI: 10.1007/s10772-020-09672-4
18. Librosa.org [Электронный ресурс]. URL: <https://librosa.org/doc/0.10.2/feature.html> (дата обращения: 17.03.2024)
19. Rainio O., Teuvo J., Klén R. Evaluation metrics and statistical tests for machine learning // Scientific Reports. 2024. Vol. 14, № 1. P. 6086
20. Ottoni L. T. C., Ottoni A. L. C., Cerqueira J. D. J. F. A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta-Learning // Electronics. 2023. Vol. 12, № 23. P. 4859. DOI: 10.3390/electronics12234859
21. Recognition of emotions in speech using convolutional neural networks on different datasets / M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, K. Kaczor // Electronics. 2022. Vol. 11, № 22. P. 3831
22. Ottoni L. T. C., Cerqueira J. D. J. F. Optimizing Speech Emotion Recognition: Evaluating Combinations of Databases, Data Augmentation, and Feature Extraction Methods // Proceedings of the XVI Brazilian Congress on Computational Intelligence. Salvador, Brazil. 2023. P. 8-11
23. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset / H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, R. Verma // IEEE Transactions on Affective Computing. 2014. Vol. 5. P. 377-390. DOI: 10.1109/TAFFC.2014.2336244
24. Livingstone S., Russo F. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English // PLoS ONE. 2018. Vol. 13. DOI: 10.1371/journal.pone.0196391
25. SAVEE (Surrey Audio-Visual Expressed Emotion) database [Электронный ресурс]. URL: <https://personalpages.surrey.ac.uk/p.jackson/SAVEE> (дата обращения: 16.03.2024)
26. Pichora-Fuller M. K., Dupuis K. Toronto emotional speech set (TESS) [Электронный ресурс]. URL: <https://doi.org/10.5683/SP2/E8H2MF> (дата обращения: 16.03.2024)