Федеральное государственное автономное образовательное учреждение высшего образования
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н.Э.БАУМАНА»
(МГТУ им. Н.Э. Баумана)

Распознавание эмоций по аудиосигналу

Студент: Гунько Н. М.

Преподаватель: Спасенов А.Ю.

Москва, 2025

Содержание доклада

Введение: описание предметной области и состояние проблемы

Эмоции играют ключевую роль в человеческом общении. Автоматическое распознавание эмоций (SER, Speech Emotion Recognition) становится всё более актуальной задачей в области цифровой обработки сигналов и искусственного интеллекта, особенно в таких сферах, как:

- голосовые помощники (например, Alexa, Siri),
- автоматизированные системы обслуживания клиентов,
- психоэмоциональный анализ в телемедицине,
- системы мониторинга водителей и пилотов.

Проблема заключается в высокой вариативности аудиосигналов, зависимости от пола, тембра, языка и интонационных особенностей речи. Несмотря на наличие решений, задача по-прежнему требует оптимизации архитектур, подготовки данных и повышения устойчивости моделей к шумам и межклассовому смешению.

Анализ обучающего набора данных

Для проекта использовался открытый набор данных **RAVDESS**. Название расшифровывается как Ryerson Audio-Visual Database of Emotional Speech and Song.

Объём: 1440 аудиофайлов (24 актёра, по 60 файлов на каждого)

Эмоции (8 классов):

Формат: WAV, 48kHz, длительность ~3 сек, стерео

Анализ обучающего набора данных

Предобработка аудио:

1. Приведение к моно: усреднение двух каналов, чтобы избавиться от лишней размерности и шума.
2. Спектральное представление: применяется MelSpectrogram с n_mels=32, чтобы перейти от временной области к частотной (такой формат подходит для CNN).
3. Нормализация: с помощью AmplitudeToDB() логарифмируем амплитуду – получаем спектр в децибелах.
4. Усечение/паддинг по времени: все спектрограммы приводятся к длине 800 фреймов по времени – это позволяет использовать фиксированный размер входа в нейросеть.

Результат каждого примера: тензор размерности [1, 32, 800] – один канал, 32 мел-коэффициента, 800 временных срезов.

Метод решения задачи и архитектура модели

Для решения задачи классификации эмоций по аудиосигналу была реализована **сверточная нейронная сеть** (CNN), работающая с мел-спектрограммами, интерпретируемыми как изображения. Это позволяет эффективно применять двумерные свертки для извлечения признаков из звукового сигнала.

Входные данные:

- Мел-спектрограмма размера [B, 1, 32, 800]:
  - B – размер батча,
  - 1 – один канал (моно сигнал),
  - 32 – количество мел-частотных полос,
  - 800 – временные фреймы (продолжительность звука).

Метод решения задачи и архитектура модели

Архитектура модели:

1. **Свертка #1**:
    - Conv2D: 1 входной канал → 16 выходных каналов
    - Параметры: ядро 5×5, шаг 2, паддинг 2
    - **Результат**: [B, 16, 16, 400]
2. **Активация и пулинг**:
    - ReLU
    - MaxPool2D: ядро 2×2
    - **Результат**: [B, 16, 8, 200]

Метод решения задачи и архитектура модели

Архитектура модели:

3. **Свертка #2**:
    - Conv2D: 16 → 32 каналов, ядро 3×3, шаг 1, паддинг 1
    - ReLU
    - AdaptiveAvgPool2d((1,1)): усреднение по частотно-временной области
    - **Результат**: [B, 32, 1, 1]
4. **Полносвязная часть**:
    - Flatten: [B, 32, 1, 1] → [B, 32]
    - Linear: 32 → 8 (по числу эмоций)
    - **Результат**: [B, 8] (**логиты классов**)

Метод решения задачи и архитектура модели

Гиперпараметры:

- Функция потерь: CrossEntropyLoss
- Оптимизатор: Adam (learning rate=1e-3, weight decay=1e-4)
- Эпох: 1000
- Batch size: 16
- Scheduler: ReduceLROnPlateau – уменьшает lr при отсутствии прогресса
- Early Stopping: прерывает обучение при переобучении

Почему работает:

- Мел-спектрограмма отражает структуру речи – изменение интонации, тембра, ритма
- Свертки эффективно выявляют паттерны на временно-частотной сетке
- Глобальное усреднение помогает избежать зависимости от длины фрагмента
- Полносвязная часть агрегирует признаки в решение

Проблемы и способы их решения

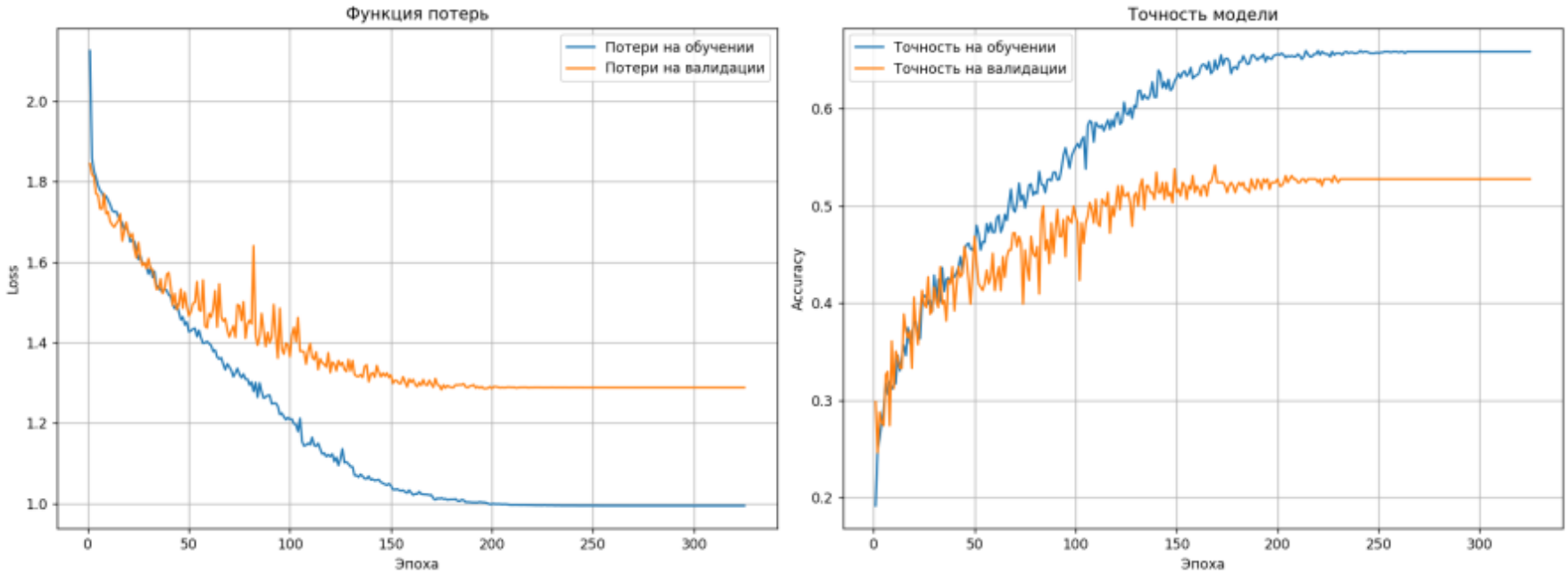
Проблемы:

- Низкая точность на сбалансированном датасете: модель училась медленно, слабо различала близкие классы (например, "calm" и "neutral")
- Переобучение: наблюдалось снижение валидационной точности после 30+ эпох
- Дисбаланс по предсказаниям: модель часто предсказывала одну и ту же эмоцию (например, angry)

Принятые меры:

- Упростили архитектуру модели для предотвращения переобучения
- Использовали AdaptiveAvgPool2D для уменьшения размерности
- Подключили ReduceLROnPlateau и EarlyStopping
- Построили confusion matrix для анализа ошибок

Демонстрация модели и оценка качества

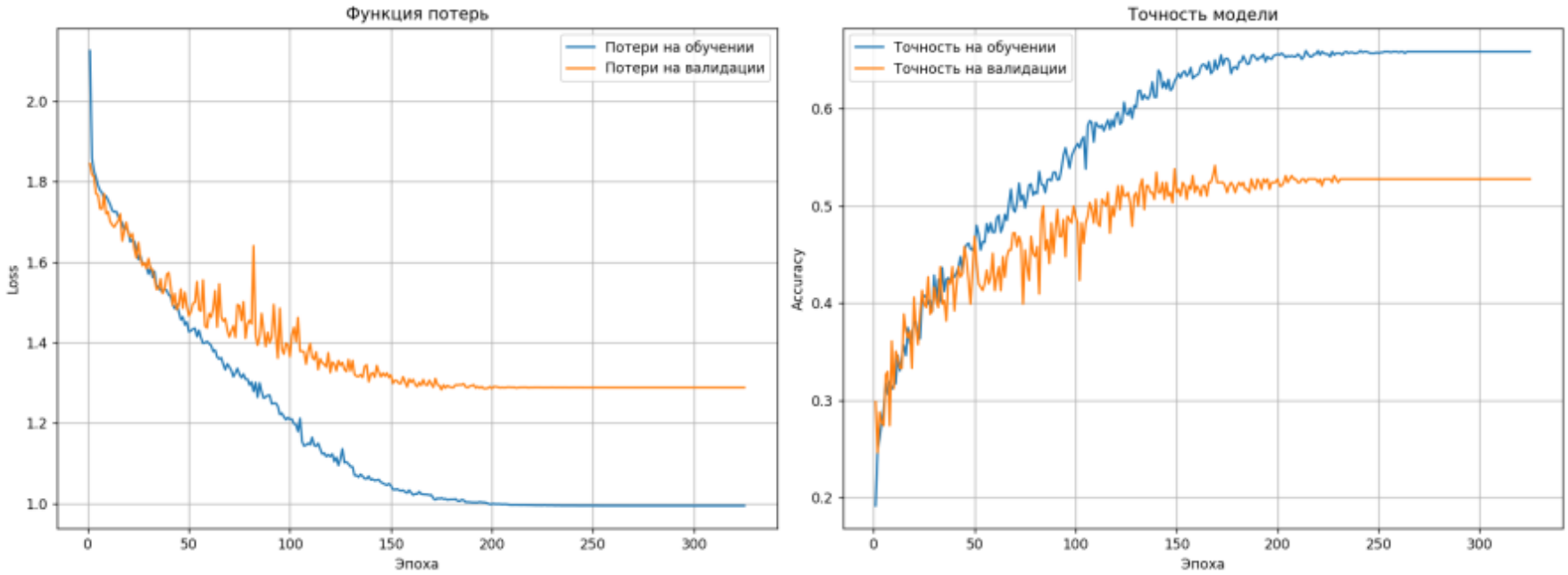


На графиках представлены кривые обучения модели по двум основным метрикам: **функция потерь** (Loss) и **точность** (Accuracy).

Слева: Функция потерь

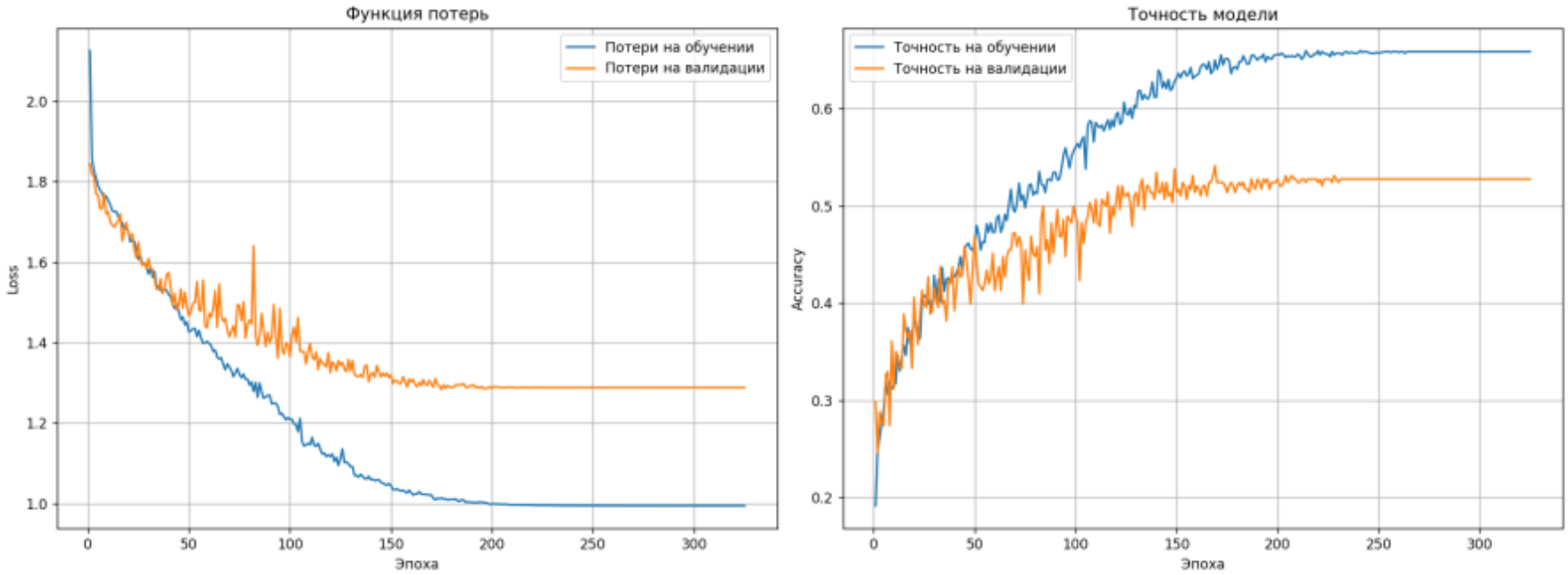
- Синяя линия – значение loss на обучающей выборке.
- Оранжевая линия – значение loss на валидационной выборке.

Демонстрация модели и оценка качества



Видно, как функция потерь на обучении стабильно убывает, тогда как на валидации она перестаёт снижаться после ~200 эпох, что сигнализирует о начале переобучения.

Демонстрация модели и оценка качества



Справа: Точность модели

- Синяя линия – точность на обучающей выборке.
- Оранжевая линия – точность на валидационной выборке.

Модель достигает ~67% точности на обучении.

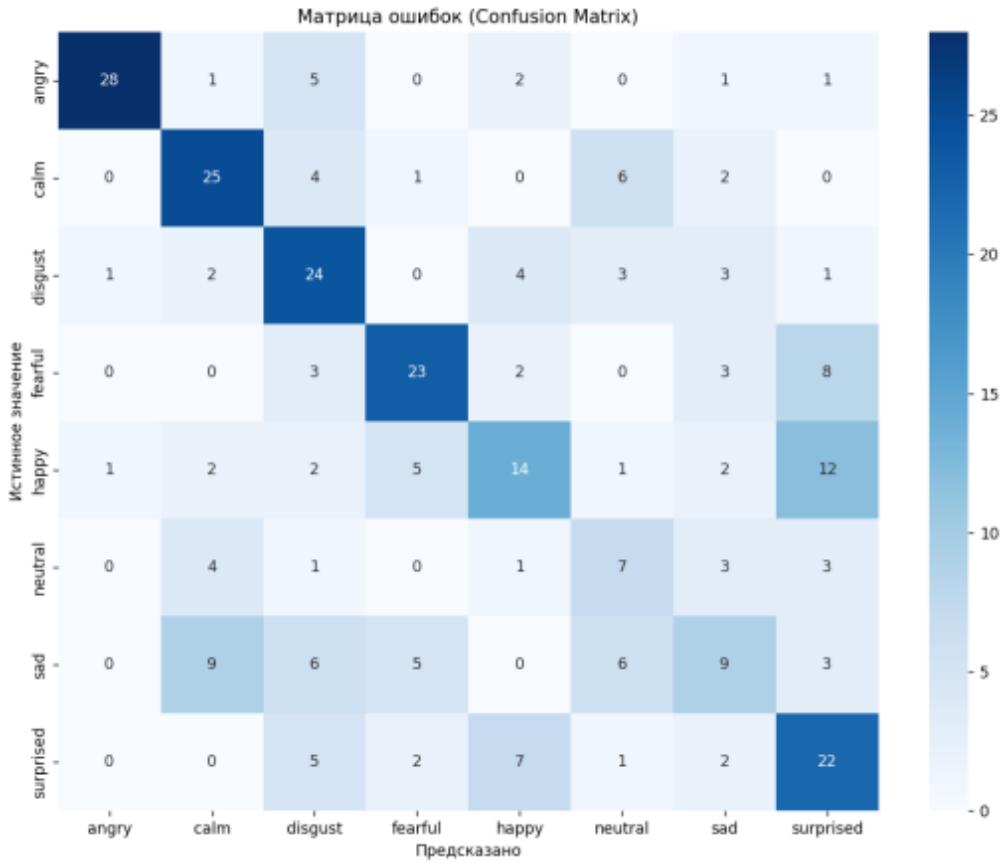
На валидации – стабильная точность ~54%.

Демонстрация модели и оценка качества

Матрица ошибок (Confusion Matrix) показывает, как модель классифицирует объекты различных классов:

- Строки – реальные классы
- Столбцы – предсказанные классы

На диагонали – количество правильно классифицированных объектов. Остальные – ошибки.

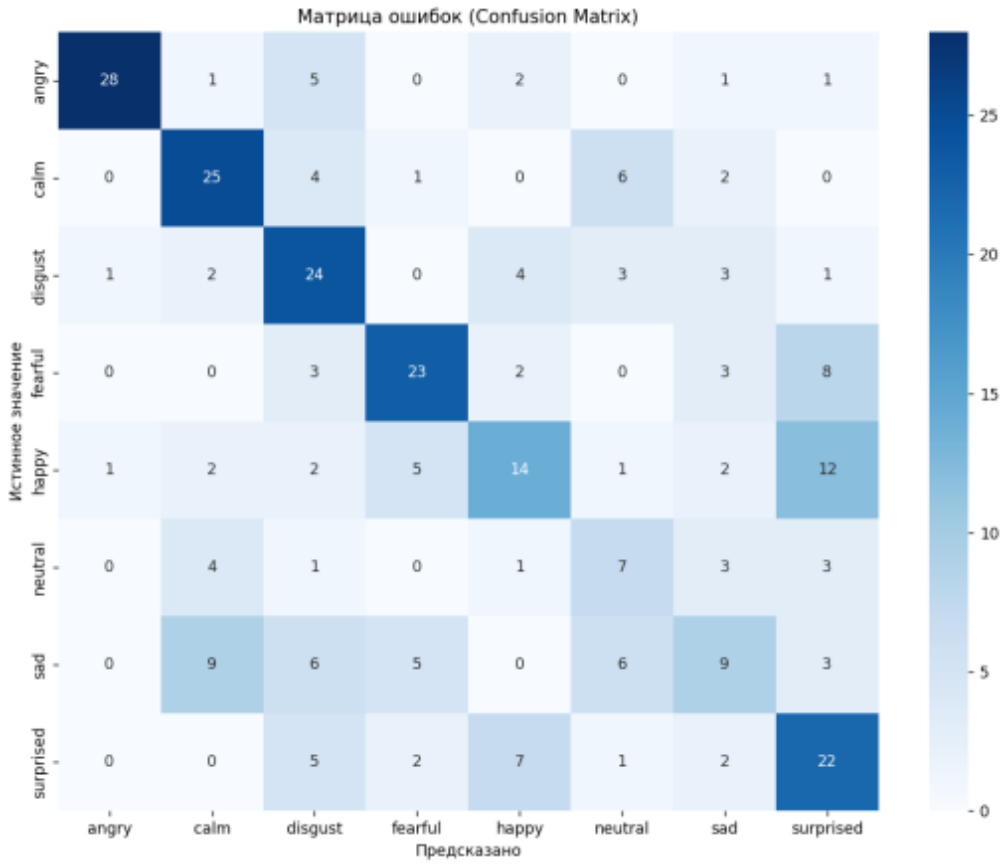


Демонстрация модели и оценка качества

Модель демонстрирует хорошее качество классификации по ряду эмоций, но также выявлены типичные ошибки:

Хорошо классифицируются:

- angry → 28 из 38 правильно (точность ~74%)
- calm → 25 из 32 (точность ~78%)
- disgust → 24 из 38 (точность ~63%)
- fearful → 23 из 41 (точность ~56%)
- surprised → 22 из 39 (точность ~56%)

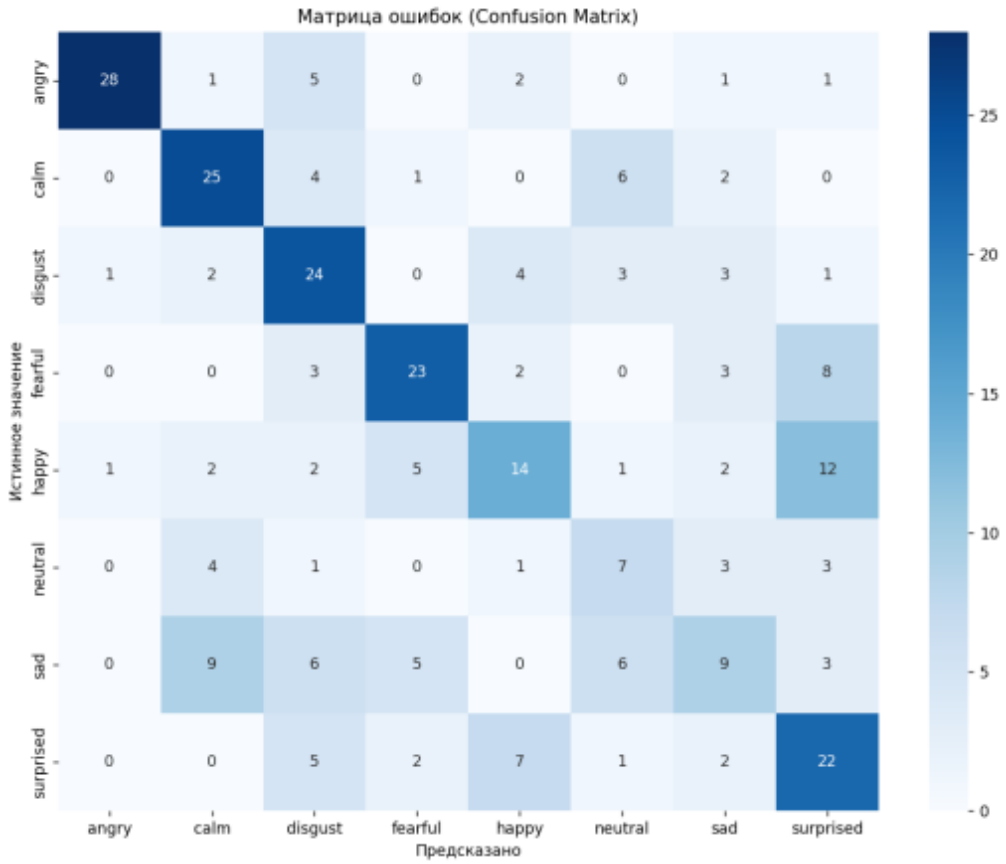


Демонстрация модели и оценка качества

Модель демонстрирует хорошее качество классификации по ряду эмоций, но также выявлены типичные ошибки:

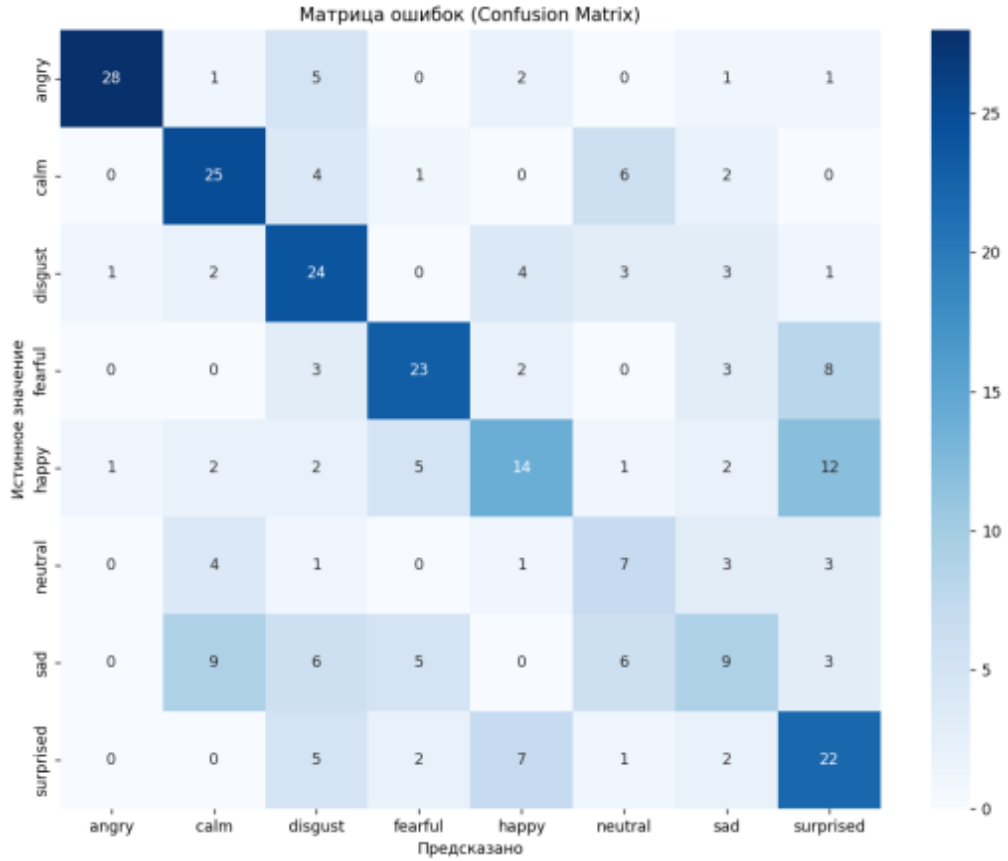
Наиболее частые ошибки:

- happy часто путается с:
  - surprised → 12 ошибок
  - fearful → 5 ошибок
- sad путается с:
  - disgust → 6 ошибок
  - fearful и calm
- neutral → расплывчато, много ошибок по всем направлениям



Демонстрация модели и оценка качества

- Эмоции, имеющие похожие акустические признаки (например, happy ↔ surprised, sad ↔ calm/disgust) часто перепутываются.
- Лучше всего различаются яркие эмоции: angry, calm, disgust.



Выводы

- Построен полный ML-пайплайн: от загрузки и преобразования аудио до визуализации результатов.
- Реализована простая и устойчивая CNN для спектрограмм, позволяющая эффективно обрабатывать аудиоэмоции.
- Модель показала достойные результаты на ограниченном датасете без сложных предобученных блоков.