УДК 004.855.5 **DOI: 10.18413/2518-1092-2022-8-3-0-5**

Жихарев А.Г.[1]
Черных В.С.[2]

КЛАССИФИКАЦИЯ РЕЧЕВЫХ ДАННЫХ ПО ЭМОЦИОНАЛЬНОМУ ФОНУ

[1)] Белгородский государственный технологический университет им. В.Г. Шухова, ул. Костюкова, 46, Белгород, 308012, Россия
[2)] Белгородский государственный университет, ул. Победы, 85, Белгород, 308015, Россия

e-mail: zhikharev@bsu.edu.ru

Аннотация
В данной работе рассматривается алгоритм классификации речевых данных по эмоциональному фону, разработанный авторами. В частности, описывается нейронная сеть, созданная с целью распознавания восьми различных эмоций в речи. Для обучения нейронной сети была использована обучающая выборка, полученная из датасета RAVDESS, который содержит 1440 аудиофайлов. Эти аудиофайлы содержат речь 24 актеров (12 женщин и 12 мужчин) с нейтральным североамериканским акцентом.
В работе описывается процесс обучения нейронной сети с использованием библиотеки Keras, включая архитектуру сети, размеры слоев, функции активации и методы оптимизации. Также обсуждаются этапы предварительной обработки и подготовки исходных аудиоданных перед обучением сети.
Полученные результаты исследования показывают, что разработанная нейронная сеть обладает высокой производительностью и способностью распознавать эмоции с точностью 80%.
Ключевые слова: аудиопризнаки; аудио; аудиофайл; аудиоданные; эмоциональный фон; классификация; модель; слой
Для цитирования: Жихарев А.Г., Черных В.С. Классификация речевых данных по эмоциональному фону // Научный результат. Информационные технологии. – Т.8, №3, 2023. – С. 34-44. DOI: 10.18413/2518-1092-2022-8-3-0-5

Zhikharev A.G.[1]
Chernykh V.S.[2]

CLASSIFICATION OF SPEECH DATA BY EMOTIONAL BACKGROUND

[1)] Belgorod State Technological University named after V.G. Shukhov, 46 Kostyukova Str., Belgorod, 308012, Russia
[2)] Belgorod State University, 85 Pobedy Str., Belgorod, 308015, Russia

e-mail: zhikharev@bsu.edu.ru

Abstract
In this paper, the algorithm of classification of speech data by emotional background, developed by the authors, is considered. In particular, it describes a neural network created to recognize eight different emotions in speech. To train the neural network, a training sample obtained from the RAVDESS dataset, which contains 1440 audio files, was used. These audio files contain the speech of 24 actors (12 women and 12 men) with a neutral North American accent.
The paper describes the process of training a neural network using the Keras library, including the network architecture, layer sizes, activation functions and optimization methods. The stages of preprocessing and preparation of the original audio data before training the network are also discussed.
The results of the study show that the developed neural network has high performance and the ability to recognize emotions with an accuracy of 80%.
Keywords: audio attributes; audio; audio file; audio data; emotional background; classification; model; layer

For citation: Zhikharev A.G., Chernykh V.S. Classification of speech data by emotional background // Research result. Information technologies. – T.8, №3, 2023. – P. 34-44. DOI: 10.18413/2518-1092-2022-8-3-0-5

ВВЕДЕНИЕ

В данный момент мы находимся в пике оптимистического развития искусственного интеллекта, и всё чаще слышим о создании новых нейронных сетей для решения самых разнообразных и важных проблем, начиная с постановки медицинских диагнозов и заканчивая усовершенствованием цифровых помощников [1]. В современном мире обработка и анализ больших объемов данных становятся все более важными задачами. Особое внимание уделяется анализу речевых данных, которые могут содержать важную информацию о настроении и эмоциональном состоянии говорящего. Одним из ключевых направлений в этой области является классификация речевых данных по эмоциональному фону.

В данной работе будет рассмотрен модуль для обработки голосовых данных и распознавания эмоций.

НАЧАЛЬНЫЕ УСЛОВИЯ

Выбор обучающих данных является важнейшей составляющей в обучении нейронной сети, поэтому тут необходимо выбрать наикачественнейшие данные, выбор пал на датасет RAVDESS, он обоснован тем, что аудиофайлы записаны на высококачественном оборудовании. Только аудио часть датасета RAVDESS содержит 1440 файлов (.wav) с речью 24 актеров (12 женщин, 12 мужчин) на нейтральном североамериканском акценте. Так же сами аудио файлы имеют удобные для понимания и расшифровки метки, пример на рисунке.

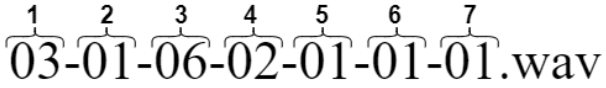

Рис. 1. Пример названия аудиофайла
Fig. 1. Example of an audio file name

1. Режим (01 = видео с аудиоряд, 02 = только видео, 03 = только аудио.
2. Вокальный канал (01 = речь, 02 = песня).
3. Эмоция (01 = нейтрально, 02 = спокойно, 03 = счастье, 04 = грусть, 05 = злость, 06 = испуг, 07 = отвращение, 08 = удивление).
4. Эмоциональная напряженность (01 = нормальная, 02 = сильная). ПРИМЕЧАНИЕ: Для "нейтральной" эмоции не существует сильной интенсивности.
5. Фраза (01 = "Дети разговаривают у двери", 02 = "Собаки сидят у двери").
6. Повторение (01 = 1-е повторение, 02 = 2-е повторение).
7. Актер (с 01 по 24. Актеры с нечетными номерами – мужчины, актеры с четными номерами – женщины).

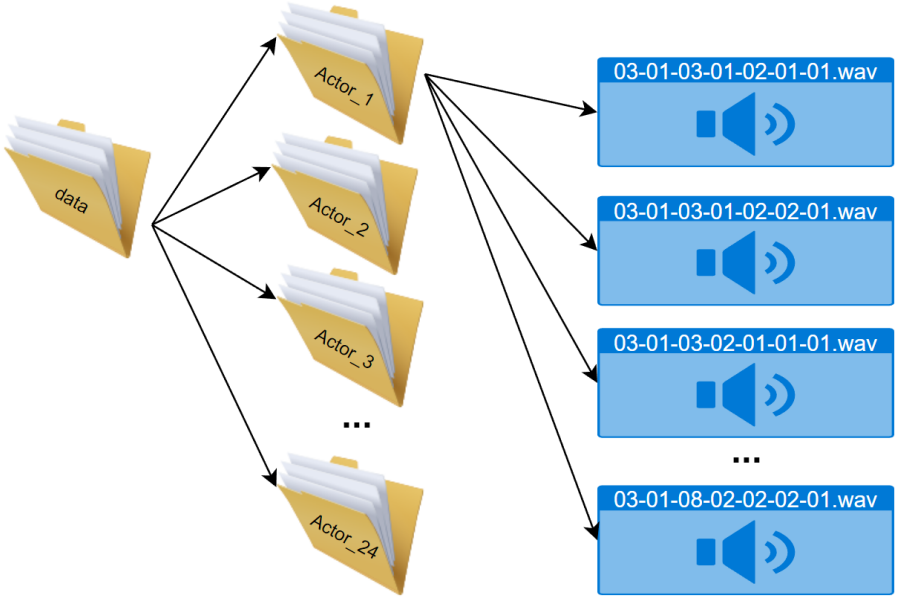

Рис. 2. Структура хранения аудиофайлов
Fig. 2. Structure of audio file storage

АУДИОПРИЗНАКИ

Эмпирическим путём было выявлено, что для получения хороших результатов подходят следующие аудиопризнаки:

1. MFCC (Mel-Frequency Cepstral Coefficients Мел-кепстральные коэффициенты) – это метод извлечения характеристик из аудиосигнала, который используется для представления звуковых сигналов в виде векторов признаков. Он является одним из наиболее популярных методов в анализе и распознавании речи.

2. Chroma (хрома) в контексте аудио-обработки является вектором, который представляет распределение энергии звукового сигнала по музыкальным тональностям (нотам). Chroma-вектор позволяет оценить наличие и относительную интенсивность различных музыкальных тональностей в аудиозаписи [2].

3. Мел-спектрограмма – это представление звукового сигнала, которое позволяет увидеть, какая часть частот присутствует в звуке в разные моменты времени. Она показывает, как энергия звука распределена по различным частотам в зависимости от времени [3].

4. RMS (Root Mean Square Среднеквадратичный корень) – это статистическая мера, используемая для измерения среднеквадратического значения амплитуды звукового сигнала. Она представляет собой квадратный корень из среднего значения квадратов амплитудных значений во временной области сигнала.

ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ

Извлечение аудио признаков происходит следующим образом, берётся аудиофайл для дальнейшего анализа, затем происходит сегментация аудио файла, это сделано для того, чтобы получить одинаковое количество признаков из аудиофайлов разной продолжительности, после этого происходит получение аудио признаков, дальше происходит их нормализация, затем аудио признаки объединяются и в итоге добавляются в общий массив признаков других аудиофайлов.

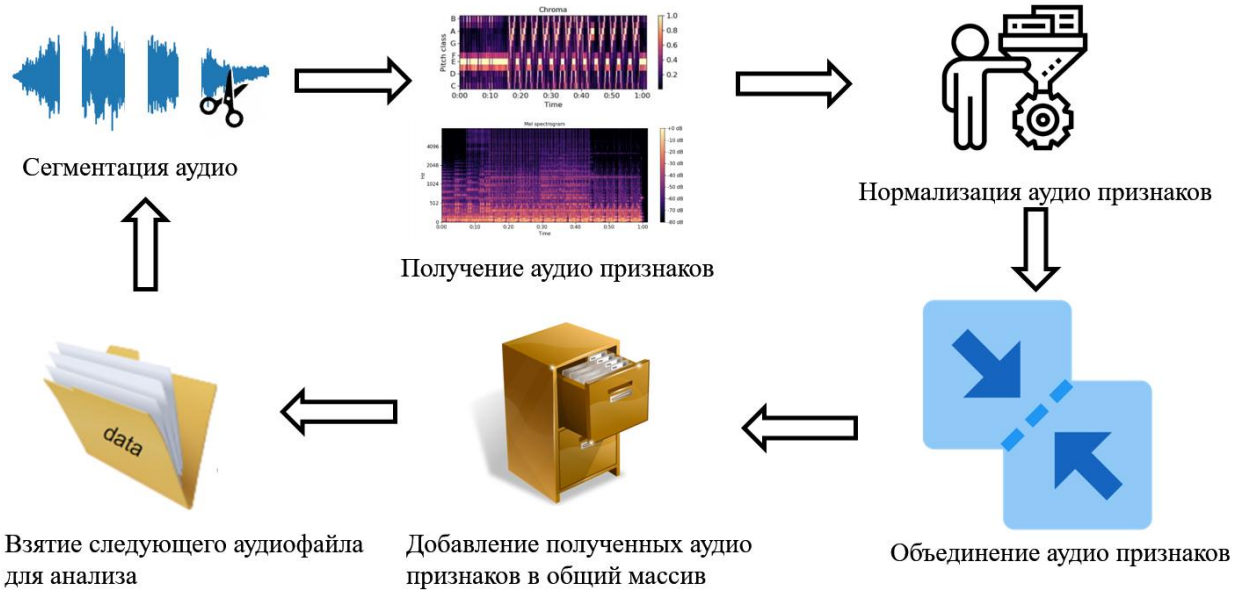

Рис. 3. Общая схема извлечения признаков из аудиоданных
Fig. 3. General scheme of feature extraction from audio data

Для эффективного распознавания эмоций в речевых данных важно извлечь соответствующие признаки из аудиофайлов. Однако, аудиофайлы могут иметь разную длительность, что создает проблемы при получении универсальных и сравнимых признаков для анализа. В целях решения этой проблемы предлагается следующий подход:

1. Округление длительности – В первую очередь, каждая длительность аудиофайла округляется до ближайшей целой секунды.

2. Разбиение на сегменты – Затем каждый аудиофайл разбивается на 20 равных сегментов. Это позволяет получить более детальное представление аудио-сигнала и учитывать его изменения во времени. Каждый сегмент имеет одинаковую длительность, что обеспечивает сравнимость признаков между различными аудиофайлами.

3. Извлечение признаков – После разбиения на сегменты из каждого сегмента извлекаются признаки, извлекаются они библиотекой librosa. Эти признаки будут включать такие характеристики, как MFCC, хромаграммы, мел-спектрограммы и RMS. Извлечение признаков позволяет представить аудиофайлы в виде числовых векторов, содержащих информацию о специфических акустических характеристиках каждого сегмента.

Если аудиофайлы имеют длительность от 3 до 5 секунд, то деление на 20 частей имеет смысл, чтобы получить более мелкие сегменты и извлечь признаки из них.

Каждый аудиофайл разбивается на 20 равных сегментов. При этом каждый сегмент будет иметь длительность, соответствующую 1/20 от общей длительности аудиофайла. Таким образом, каждый сегмент будет иметь продолжительность от 0.15 до 0.25 секунды. Важно отметить, что получаемые массивы данных из RMS строго зависят от длинны аудиозаписи, поэтому массивы имеют фиксированный размер, если данных недостаточно, то они дополняются нулями, иначе обрезаются.

Нормализация данных в аудио признаках выполняется с целью привести их к общему масштабу и диапазону значений. Опытным путём было выявлено, что метод нормализации L2-нормы, которая основывается на вычислении евклидовой нормы векторов признаков, имеет наивысшую эффективность. Этот подход позволяет привести значения признаков к единичной длине, что упрощает их сравнение и улучшает стабильность процесса обучения.

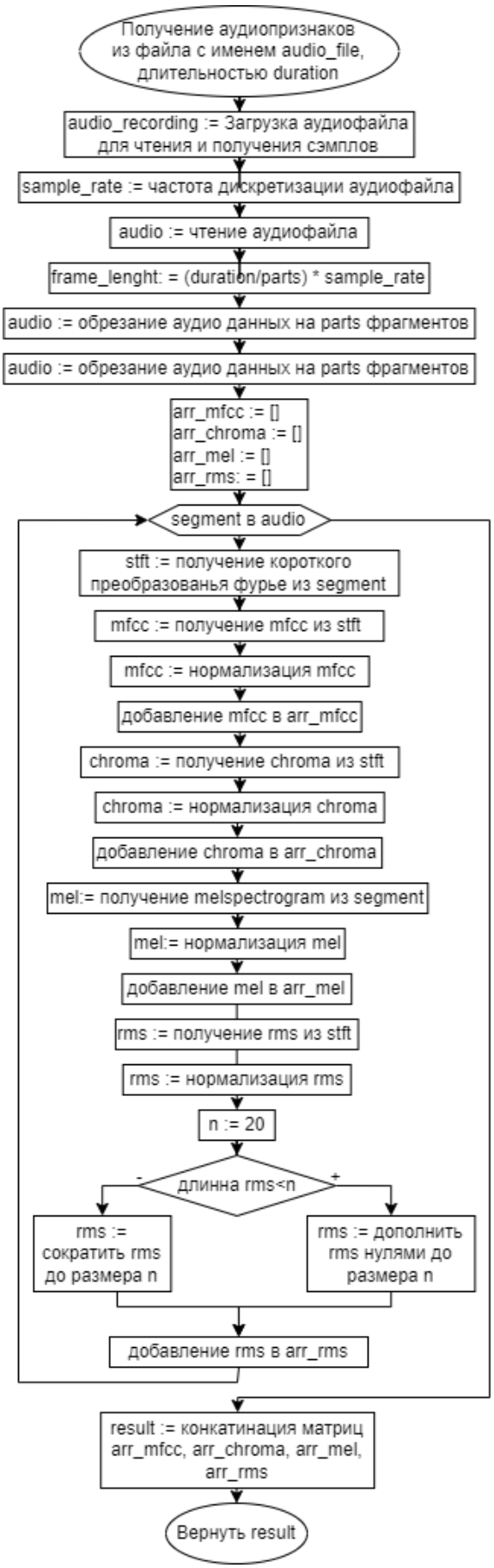

Рис. 4. Получение аудио признаков из файла
Fig. 4. Getting audio features from a file

МОДЕЛЬ НЕЙРОННОЙ СЕТИ

BatchNormalization: Этот слой нормализует входные данные путем стандартизации их по среднему значению и дисперсии. Это помогает улучшить стабильность и скорость обучения модели.

Bidirectional LSTM: LSTM (Long Short-Term Memory сети долгой краткосрочной памяти) – это рекуррентный слой, который способен сохранять информацию о предыдущих состояниях. Bidirectional LSTM обрабатывает последовательности в обоих направлениях, что позволяет модели улавливать и использовать информацию из прошлого и будущего контекста [4]. Он эффективно моделирует зависимости во временных данных, таких как звуковые сигналы речи.

LSTM: Еще один слой LSTM используется для дальнейшего извлечения временных зависимостей из последовательностей данных. Это помогает модели уловить более сложные и долгосрочные зависимости в аудиоданных [5].

Conv1D: Слой одномерной свертки используется для обнаружения локальных шаблонов и особенностей в аудиоданных.

Dropout: Слой Dropout используется для предотвращения переобучения модели путем случайного обнуления некоторых элементов выходных данных во время обучения. Это помогает улучшить обобщающую способность модели и предотвратить переобучение.

Flatten: Слой Flatten преобразует многомерный выход из предыдущего слоя в одномерный вектор. Он подготавливает данные для передачи в полносвязные слои модели.

Dense: Полносвязные слои применяются для классификации. Они содержат нейроны, соединенные со всеми нейронами предыдущего слоя.

Совмещение Conv1D и LSTM позволяет моделировать как глобальные, так и локальные зависимости в данных. LSTM слои помогают захватывать долгосрочные зависимости и общий контекст, в то время как Conv1D слои улавливают локальные паттерны и временные особенности данных. Комбинирование этих двух типов слоев дает возможность модели извлекать более информативные признаки из входных данных и лучше понимать структуру временных рядов.

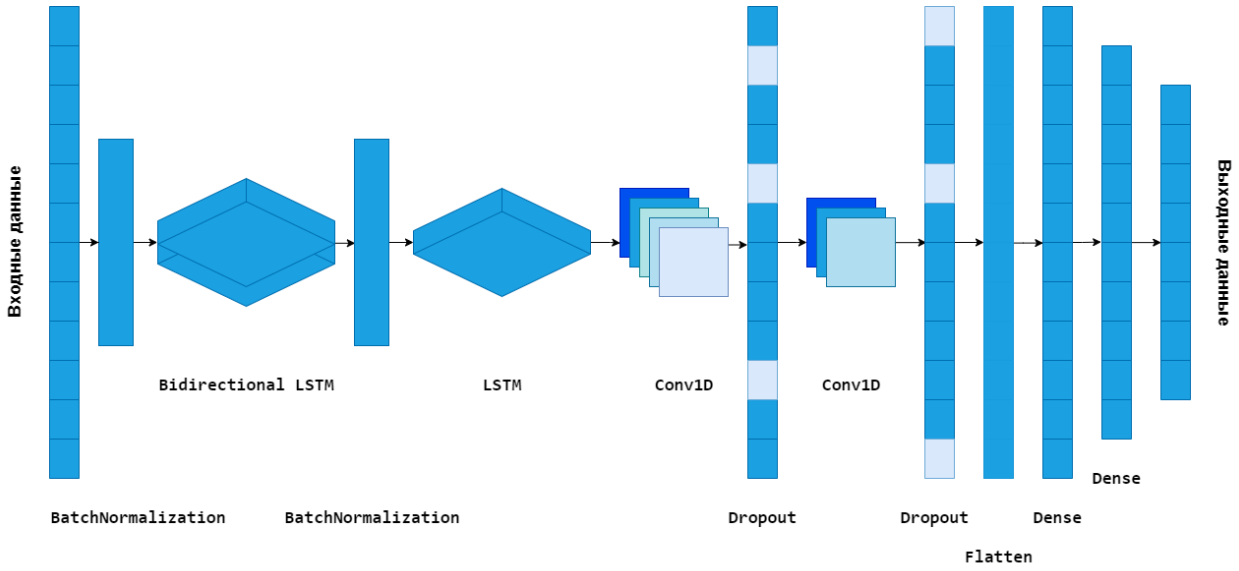

Рис. 5. Структура модели нейронной сети
Fig. 5. Structure of the neural network model

Первый слой BatchNormalization имеет значение momentum=0.91 означает, что модель будет сохранять 91% информации о статистике предыдущего пакета и использовать его для нормализации текущего пакета данных. Это позволяет модели быстрее адаптироваться к изменениям в данных и повышает устойчивость обучения. Более высокое значение momentum может привести к более

стабильной и более медленной адаптации, в то время как более низкое значение может привести к более быстрой, но менее стабильной адаптации.

Второй слой Bidirectional LSTM имеет 240 нейронов и функцию активации 'tanh', все слои имеют данную функцию активации, так как она показала наивысшую эффективность.

Третий слой BatchNormalization имеет значение momentum=0.95.

Четвертый слой LSTM имеет 140 нейронов.

Пятый слой Conv1D имеет 36 нейронов с ядром свёртки 5 и L2-регулиризация со значением 0,0001(kernel_regularizer=regularizers.l2(0.0001)).

Шестой слой Dropout имеет значение 0.2.

Седьмой слой Conv1D имеет 20 нейронов с ядром свёртки 3.

Восьмой слой Dropout имеет значение 0.2.

Десятый слой Dense имеет 140 нейронов.

Одиннадцатый слой Dense имеет 40 нейронов.

Двенадцатый и он же выходной слой имеет 8 нейронов, т.е. равную количество эмоций, а так же функцию активации ‘softmax’.

В данной модели используется функция потерь categorical_crossentropy, так как решается задача многоклассовой классификации.

Оптимизатор adam выбран для эффективного обновления весов модели в процессе обучения, обеспечивая быструю сходимость и хорошие результаты.

РЕЗУЛЬТАТЫ РАБОТЫ

Исходные данные были поделены на обучающие и тестовые, в тестовые данные входят 20% от изначальной.

Рис. 6. Аудиоданные из датасета RAVDESS
Fig. 6. Audio data from the RAVDESS dataset

Рис. 7. Обучающие аудиоданные из датасета RAVDESS
Fig. 7. Training audio data from the RAVDESS dataset

Рис. 8. Тестовые аудиоданные из датасета RAVDESS
Fig. 8. Test audio data from the RAVDESS dataset

Каждый класс имеет различное количество аудио файлов как в обучающем, так и в тестовом наборе данных. Это позволяет модели получить разнообразный набор примеров и обучиться на различных эмоциональных фоновых состояниях.

В результате обучения модели на разнообразном наборе данных, получили хорошие результаты точности и потерь.

Анализируя графики, можно заметить, что уже на 20-й эпохе обучения точность модели на обучающих данных достигла значения 100%, тогда как на тестовых данных она составила 78%. Это свидетельствует о том, что модель успешно обучается и справляется с классификацией эмоций на обучающих и незнакомых данных.

Кроме того, потери на обучающих данных приближаются к нулю, что указывает на эффективность обучения модели на тренировочном наборе данных. В то же время, потери на тестовых данных составляют примерно 0.77, что говорит о том, что модель делает некоторые ошибки в предсказаниях на незнакомых данных, но все равно достигает хороших результатов.

Однако, самые лучше показатели были достигнуты на 67-й эпохе. Точность модели увеличилась до 83%, что является высоким значением для данной задачи. При этом потери также немного увеличились до 0.81, но оставались на приемлемом уровне.

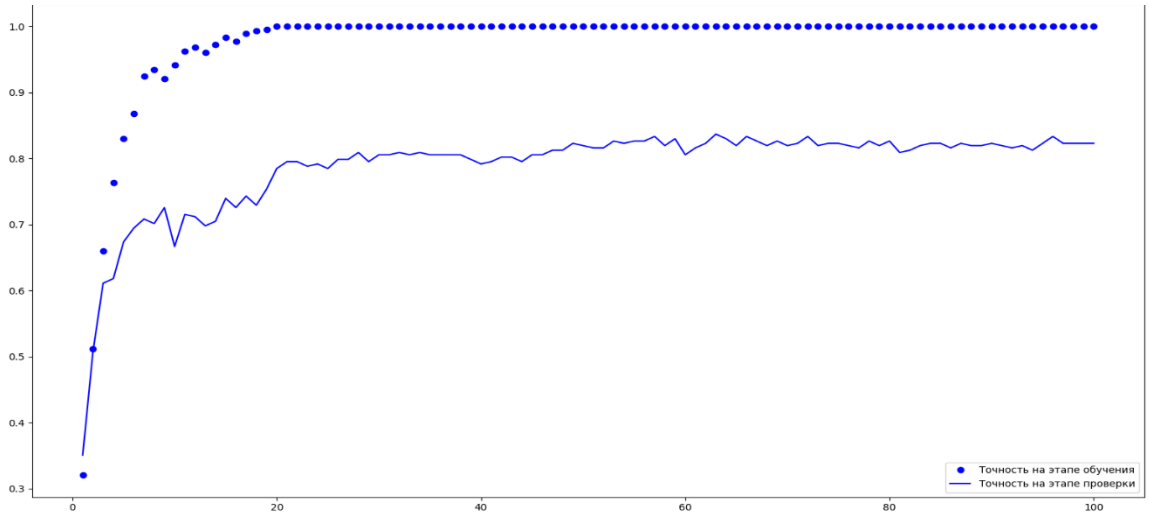

Рис. 9. Точность на этапах обучения и проверки
Fig. 9. Accuracy at the stages of training and verification

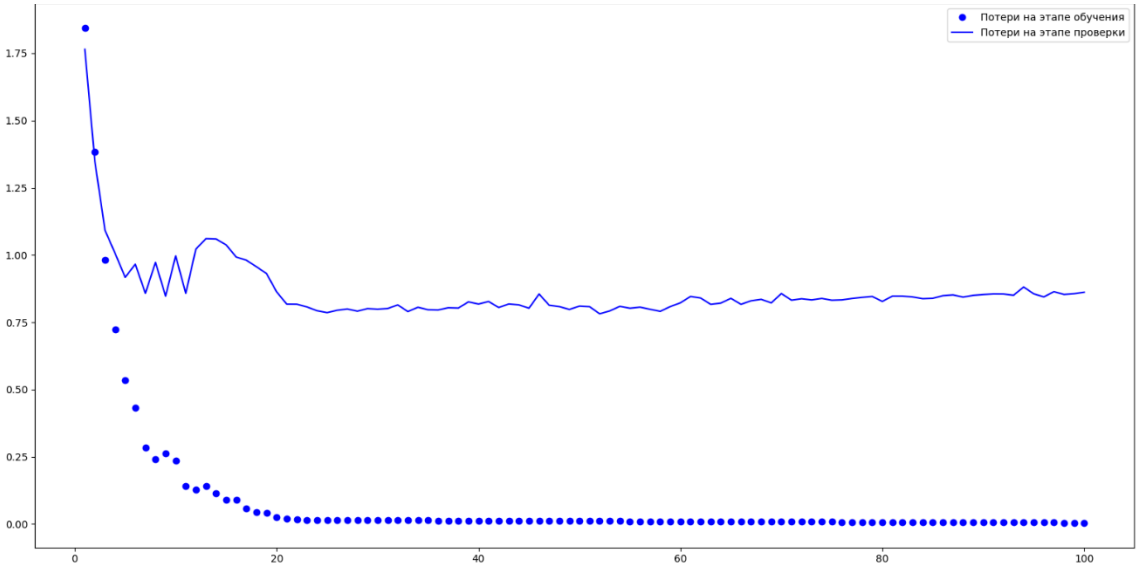

Рис. 10. Потери на этапах обучения и проверки
Fig. 10. Losses at the stages of training and verification

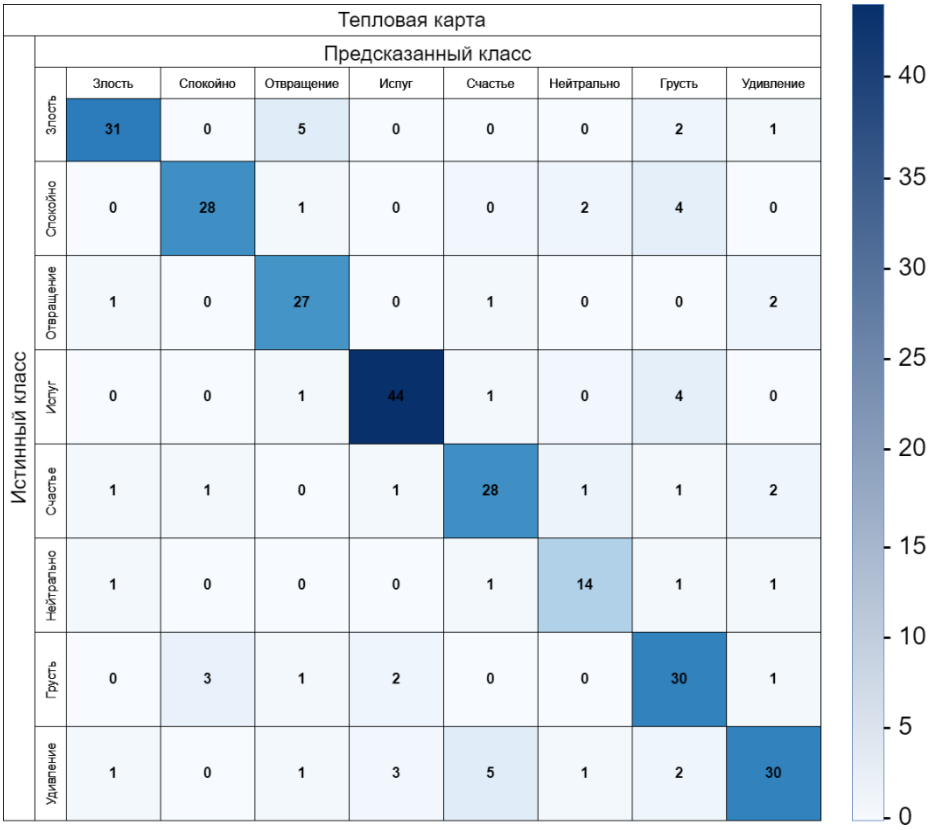

Рис. 11. Тепловая карта результатов обученной модели
Fig. 11. Heat map of the results of the trained model

ЗАКЛЮЧЕНИЕ

С использованием библиотеки Keras и соответствующих алгоритмов была создана модель, которая показала высокую точность и низкие потери. Графики точности и потерь подтверждают успешность обучения модели, достигнув наиболее оптимальных показателей на 67 эпохе с точностью в 83% и потерями в 0.81.

Для дальнейшего улучшения модели классификации эмоций на основе аудиоданных можно увеличить размер обучающего набора данных, чтобы обеспечить большую обобщающую способность модели. Несмотря на достигнутые высокие показатели точности и потерь, данная модель имеет потенциал для дальнейшей оптимизации и улучшения. Дальнейшие исследования и разработки в этой области могут привести к созданию более точных и эффективных моделей классификации эмоций на основе аудиоданных.

Список литературы

1. Шолле Ф. Глубокое обучение на Python. 2-е межд. издание. – СПб.: Питер, 2023. – 576 с. – ISBN 978-5-4461-1909-7.

2. Han K., Lee K., Kim H.G. Music emotion recognition using chroma feature-based probabilistic neural network. Multimedia Tools and Applications. – 2017. – Том №76, Выпуск №3. – С. 3691-3710.

3. Getting to Know the Mel-Spectrogram. [Электронный ресурс] – Электрон, дан., 2019. – URL: https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0

4. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Nature. – 2005. – Том №18, Выпуск №5-6. – С. 602-610.

5. Understanding LSTM Networks. [Электронный ресурс] – Электрон, дан., 2015. – URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

References

1. Chollet F. Deep Learning with Python. 2nd interd. edition. – SPb.: St. Petersburg: St. Petersburg. – 576 p. – ISBN 978-5-4461-1909-7.
2. Han K., Lee K., Kim H.G. Music emotion recognition using chroma feature-based probabilistic neural network. Multimedia Tools and Applications. – 2017. – V. №76, Issue №3. – P. 3691-3710.
3. Getting to Know the Mel-Spectrogram. [Electronic resource] – Electronic data, 2019. – URL: https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0.
4. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Nature. – 2005. – V. №18, Issue №5-6. – P. 602-610.
5. Understanding LSTM Networks. [Electronic resource] – Electronic data, 2015. – URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Жихарев Александр Геннадиевич, доктор технических наук, доцент, доцент кафедры программного обеспечения вычислительной техники и автоматизированных систем
Черных Владимир Сергеевич, студент 1 курса Магистратуры, направления «Прикладная информатика», кафедра прикладной информатики и информационных технологий, институт инженерных и цифровых технологий

Zhikharev Alexander Gennadievich, Doctor of Technical Sciences, Associate Professor, Associate Professor of the Department of Software for Computer Engineering and Automated Systems
Chernykh Vladimir Sergeevich, 1st year Master's student, Applied Informatics, Department of Applied Informatics and Information Technologies, Institute of Engineering and Digital Technologies