*УДК 004.94* *1.2.2*

# ПРИМЕНЕНИЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ РЕЧИ

***Д.И. Карпенкова, А.С. Катасёв***

Казанский национальный исследовательский технический университет
им. А.Н. Туполева-КАИ
Российская Федерация, 420111, г. Казань, ул. К. Маркса, д. 10

**Аннотация.** В работе рассмотрено решение задачи распознавания эмоциональной окраски речи на основе построения и исследования нейросетевой модели. Проанализированы типовые методы классификации эмоций. Для решения задачи обоснована целесообразность использования категориальной модели представления эмоций как наиболее эффективной. В качестве объекта исследований выступают аудиозаписи человеческой речи. Для анализа значений параметров аудиозаписей, таких как мел-кепстральные коэффициенты, спектрограммы и хроматограммы, предложено использовать нейросетевую модель. В качестве исходных данных для анализа и нейросетевого моделирования использовано несколько наборов англоязычных аудиоданных, найденных на платформе kaggle. Исходный набор данных выделяет семь классов (эмоций): счастье, удивление, нейтральная эмоция, гнев, печаль, страх, отвращение. Общее число аудиозаписей в сформированном наборе составляет 48648. Исходные данные были представлены в виде аудиозаписей различной длины. Для обучения нейросетевой модели из аудиозаписей были извлечены характерные признаки и проведена аугментация. По исходным данным рассчитаны значения 162 параметров аудиозаписей с получением единой таблицы данных для анализа. Описан процесс подготовки данных к анализу и моделированию. Проведено разбиение данных на обучающее и тестовое множества, а также построение и исследование нейросетевой модели в виде сверточной нейронной сети. Для оценки эффективности построенной модели произведена оценка точности, полноты и F-меры построенной модели. Результаты исследований показали, что построенная модель является достаточно эффективной и может быть использована в составе интеллектуальной системы поддержки принятия решений.

**Ключевые слова:** нейросетевая модель, эмоциональная окраска речи, анализ аудиоданных, моделирование.

## Введение

На сегодняшний день все больше компаний задумывается об автоматизации работы службы поддержки путем внедрения различных helpdesk-решений. Исследование [1] показывает, что большая часть организаций (77 % исследованных компаний) все еще указывают телефон в качестве способа для связи. Соответственно, существует необходимость обработки входящих заявок, поступающих по этому каналу связи.

Как известно, обработка аудиозаписей, а в особенности обработка человеческой речи – задача, решению которой посвящены многие работы [2-5]. Одной из областей, в которой развивается распознавание речи, является распознавание эмоциональной окраски речи говорящего.

Под эмоциями понимаются особые психические процессы, которые выражают реакцию индивида на влияние внутренних или внешних раздражителей, имеющие ярко выраженный субъективный оттенок в виде непосредственных переживаний [6]. Для регуляции жизни человека эмоции являются важнейшим фактором, включающим в себя все аспекты чувствительности. На практике не существует однозначного определения или метода измерения эмоций, что приводит к различным подходам к процессу их классификации [7].

На сегодняшний день принято использовать одну из двух формальных моделей представления эмоций: категориальную (дискретную) или многомерную (непрерывную) [8].

Дискретный подход предполагает существование первичных эмоциональных состояний, которые также называют базовыми, определение которых также не является однозначным – в различных исследованиях выделяют от шести до 22 видов эмоциональных состояний. В непрерывном подходе эмоция рассматривается в качестве базиса в многомерном координатном пространстве. Изменение величины, присутствующей в определенном измерении, характеризует изменение эмоционального состояния и его интенсивности. Кроме того, существует гибридный подход, который предполагает комбинацию дискретного и многомерного. Все эти подходы основываются на общем допущении, что в один момент времени может быть выражена только одна эмоция.

В данной работе будет использован дискретный подход определения эмоций. Рассмотрим типовые методы классификации аудиоданных.

## Методы классификации аудиоданных

В последние годы аффинные вычисления становятся предметом интереса множества ученых [9]. Для обеспечения эффективного взаимодействия между компьютером и человеком создаются интеллектуальные системы, способные обрабатывать, распознавать и интерпретировать эмоциональное состояние человека, чтобы затем адаптировать свое поведение соответствующим образом.

Исследование применимости нейронных сетей является новой областью, которая основана на концепции мягких вычислений [10-15]. Наравне с классификацией изображений в качестве задачи машинного обучения [16, 17] поднимается вопрос о применимости подобных подходов в других областях, например, классификации аудиозаписей.

В качестве входа в задаче классификации аудиозаписей обычно используется аудиофайл, содержащий отдельные звуки, шумы различной длины, музыку, речь; реже – текст или видео. Выходом является метка класса, например, при классификации аудиозаписей по полу говорящего аудиозапись может быть отнесена к классу «мужчина» или к классу «женщина».

Сама по себе проблема классификации аудиосигналов уже неоднократно затрагивалась научным сообществом в различных исследованиях, которые посвящены, например, аугментации данных классификации звуков среды, распознаванию акустических событий с использованием глубоких нейронных сетей [18, 19], классификации звуков среды при помощи сверточных нейронных сетей и т.д. Так или иначе, все эти исследования имеют общую архитектуру, основанную на глубокой сверточной или рекуррентной нейронной сети [20-22]. Каждый слой такой сети является рекуррентным, то есть получает скрытое состояние предыдущего слоя в качестве входных данных. Эта архитектура позволяет выполнять иерархическую обработку сложных временных задач и более естественно фиксировать структуру временных рядов [23]. Такие сети в различных интерпретациях оказались эффективными для решения таких задач, как распознавание речи [24, 25].

При обработке аудиозаписей можно выделить два основных шага. Первый шаг заключается в предварительной обработке аудиозаписей, которая может заключаться в нормализации амплитуды, кадрировании, блокировке кадров и оконном управлении. Второй шаг заключается в извлечении признаков, которые выделяются из каждого кадра отдельно, чтобы представить аудиоданные в виде упрощенного набора акустических признаков, извлекаемых при помощи различных методов. Обычно из каждой аудиозаписи выбирается фиксированное число наиболее энергоемких (громких) кадров (для всех рассматриваемых аудиозаписей это число одинаково), остальные кадры отбрасываются и не участвуют в рассмотрении и извлечении признаков. При рассмотрении каждого из кадров также рассматриваются и соседние (левый и правый).

Рассмотрим данный метод более подробно.

## Анализ характеристик аудиозаписей как эффективный метод распознавания эмоциональной окраски речи

В качестве исходных данных для анализа аудиофайлы представляются в виде частотной спектрограммы, сохраняемой как изображение (мел-спектрограммы), которая получается из импульсно кодовой модуляции цифровых файлов. Такое изображение обрабатывается нейронной сетью. На рис. 1-4 представлены примеры графического представления аудиосигналов, которые содержат фразу «Kids are talking by the door», произнесенную с различной эмоциональной окраской.

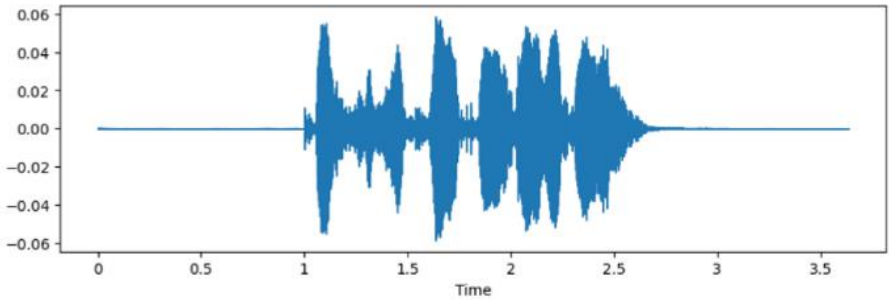


Рис. 1. Представление аудиосигнала как амплитуды по времени для эмоции «страх»

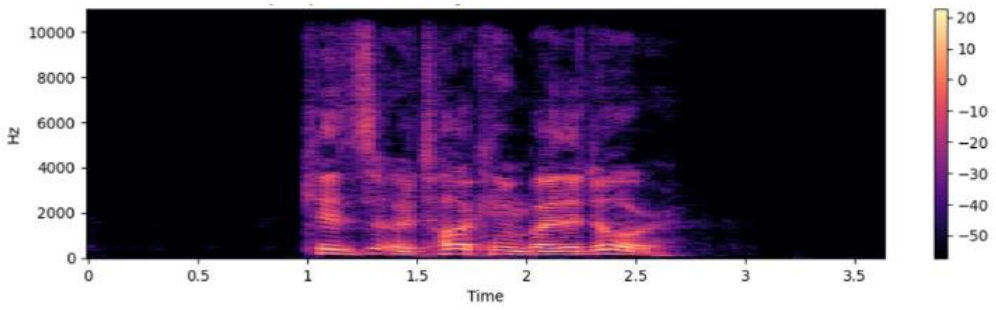


Рис. 2. Спектрограмма аудиосигнала для эмоции «страх»

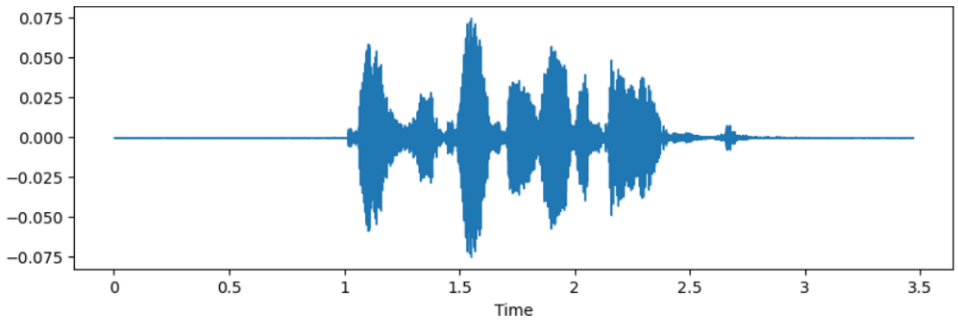


Рис.3. Представление аудиосигнала как амплитуды по времени для эмоции «счастье»

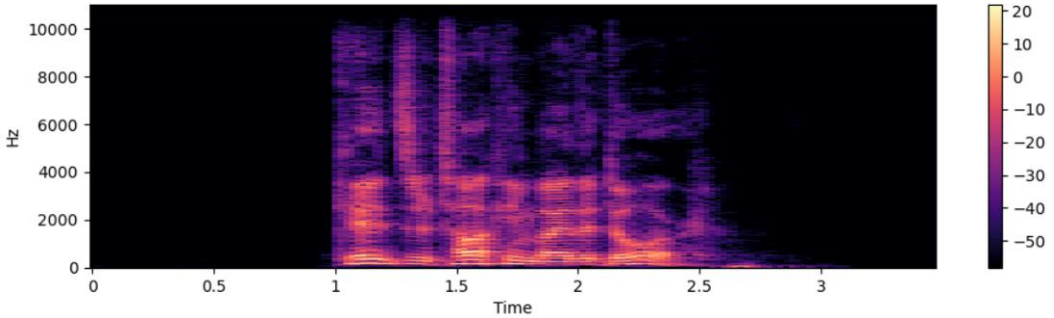


Рис. 4. Спектрограмма аудиосигнала для эмоции «счастье»

Очевидно, что для разных фраз и отдельных слов графики будут разными, поэтому такой метод обработки аудиоданных является применимым.

**Параметры аудиозаписей для определения эмоциональной окраски речи**

Далее необходимо извлечь характеристики аудиозаписей. Извлечение характеристик – очень важная часть анализа и поиска взаимосвязей между разными вещами. Данные, предоставленные в аудиоформате, не могут быть поняты моделями напрямую, необходимо преобразовать их в понятный формат, для которого используется извлечение признаков.

Звуковой сигнал представляет собой трехразмерный сигнал, в котором оси соответствуют времени, амплитуде и частоте сигнала. С помощью частоты дискретизации и примерных данных можно выполнить несколько преобразований, чтобы извлечь из него значащие характеристики.

В данной работе для построения нейросетевой модели из аудиозаписей извлекались следующие характеристики:

1. Характеристика Zero Crossing Rate – частота пересечения нуля – частота изменения знака сигнала, то есть это частота, с которой сигнал меняется с положительного на отрицательный и обратно. Данная характеристика часто используется для распознавания речи и извлечения информации о ней. Она измеряется в Гц и извлекается при помощи функции librosa.feature.zero_crossing_rate.

2. Характеристика Chroma_stft названа по методу извлечения признаков, используемому при обработке аудиосигнала для представления высоты тона и содержания гармоник аудиосигнала. Он основан на кратковременном преобразовании Фурье (STFT) и создает хромаграмму, которая представляет собой двумерное представление энергии каждого элемента цветности в каждом кадре STFT, измеряется в нормализованной энергии для каждого бита цветности в каждом кадре. Данная характеристика извлекалась при помощи функции librosa.feature.chroma_stft.

3. Характеристика MFCC – кепстральные коэффициенты мел-частоты, которые представляют собой метод выделения признаков, используемый при обработке аудиосигнала для представления спектральных характеристик аудиосигнала. Характеристика измеряется в коэффициентах и извлекается при помощи функции librosa.feature.mfcc.

4. Характеристика RMS (root mean square) – в контексте обработки аудиосигнала среднеквадратичное значение используется для измерения мощности аудиосигнала. Данная характеристика извлекалась при помощи функции librosa.feature.rms.

5. Характеристика MelSpectogram – метод извлечения признаков, используемый при обработке аудиосигнала для представления спектрального содержимого аудиосигнала. Сама характеристика MelSpectrogram представляет собой двумерное представление спектральной плотности мощности аудиосигнала, где ось частот разделена на элементы шкалы мел. Характеристика измеряется в единицах спектральной плотности мощности и извлекается при помощи функции librosa.feature.melspectrogram.

Рассмотрим вопросы подготовки исходных данных, использованных для анализа и нейросетевого моделирования.

**Подготовка исходных данных к анализу и построению нейросетевой модели**

В качестве исходных данных для анализа и моделирования использовано несколько наборов данных, найденных на портале Kaggle: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [26, 27], Surrey Audio-Visual Expressed Emotion (SAVEE) [28], Toronto emotional speech set (TESS) [29] и Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) [30]. Исходные аудиозаписи распределены по семи классам. Распределение представлено в табл.1.

Таблица 1. Распределение аудиозаписей по классам

| Класс | Наборы данных | | | | |
|---|---|---|---|---|---|
| | RAVDESS | SAVEE | TESS | CREMA-D | **Итого** |
| Нейтральная эмоция | 288 | 120 | 400 | 1087 | 1895 |
| Счастье | 192 | 60 | 400 | 1271 | 1923 |
| Печаль | 192 | 60 | 400 | 1271 | 1923 |
| Злость | 192 | 60 | 400 | 1271 | 1923 |
| Страх | 192 | 60 | 400 | 1271 | 1923 |
| Отвращение | 192 | 60 | 400 | 1271 | 1923 |
| Удивление | 192 | 60 | 400 | 0 | 652 |
| **Итого** | 1440 | 480 | 2800 | 7442 | 12162 |

Извлечение характеристик производилось при помощи средств библиотеки librosa на Python параллельно с аугментацией данных. Сначала извлекались характеристики исходных аудиозаписей, затем аудиозапись изменялась при помощи специальных функций (зашумлялась, ускорялась, замедлялась, изменялась высота звука) и характеристики извлекались из полученных аудиозаписей. Таким образом количество записей увеличилось в 4 раза.

Получаемые результаты последовательно объединялись в единую таблицу. Каждая строка таблицы содержит 162 извлеченных значения характеристик аудиозаписей и метку класса, к которому была отнесена аудиозапись. Пример данных из полученной таблицы представлен на рис.5.

| | 0 | 1 | 2 | 3 | 4 | ... | 157 | 158 | 159 | 160 | 161 | labels |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.384833 | 0.670027 | 0.725331 | 0.757509 | 0.762004 | ... | 0.000027 | 0.000026 | 0.000031 | 0.000018 | 1.582224e-06 | neutral |
| 1 | 0.349255 | 0.783358 | 0.822154 | 0.841042 | 0.844419 | ... | 0.000149 | 0.000146 | 0.000152 | 0.000137 | 1.212063e-04 | neutral |
| 2 | 0.205540 | 0.669476 | 0.666516 | 0.694322 | 0.749312 | ... | 0.000004 | 0.000005 | 0.000003 | 0.000001 | 9.096312e-08 | neutral |
| 3 | 0.196289 | 0.700014 | 0.696008 | 0.676224 | 0.726468 | ... | 0.000003 | 0.000003 | 0.000002 | 0.000001 | 1.158633e-07 | neutral |
| 4 | 0.375725 | 0.716791 | 0.752273 | 0.773176 | 0.745175 | ... | 0.000009 | 0.000007 | 0.000008 | 0.000004 | 3.281843e-07 | neutral |

Рис. 5. Фрагмент исходных данных для анализа

Нейросетевая модель требует наличия сформированных выборок для её обучения и тестирования. Грамотное формирование обучающей выборки важно для задач машинного обучения. Для формирования обучающей и тестовой выборок применялись функции библиотеки sklearn на Python. Используя функцию train_test_split, можно разбить данные на четыре части – данные независимых переменных (характеристик аудиозаписей) и данные зависимой (целевой – метки класса) переменной, каждая из этих частей разбивается на обучающую и тестовую выборки. В результате была получена обучающая выборка из 36486 записей и тестовая выборка из 12162 записей.

## Построение нейросетевой модели

Для построения нейросетевой модели распознавания эмоциональной окраски речи использованы сверточные нейронные сети. Для построения нейросетевой модели распознавания эмоциональной окраски речи использовались средства библиотеки keras на Python. Построение нейросетевой модели осуществлялось в среде JupiterLab.

При построении нейросетевой модели важную роль играют выбранные значения гиперпараметров. Гиперпараметрами называются такие параметры в машинном обучении, значения которых используются для управления процессом обучения. К гиперпараметрам можно отнести, например, топологию или размер нейронной сети, количество скрытых слоев и количество узлов в каждом слое. Производительность построенной нейросетевой модели во много зависит от выбранных значений гиперпараметров.

В качестве начальных были выбраны следующие значения гиперпараметров:

- количество входных параметров – 162 (по количеству извлеченных из аудиозаписей характеристик);
- начальное количество эпох обучения – 100;
- размер пакета (batch_size) принят равным 16 – размер пакета определяет количество записей из исходного набора данных, которое будет распространяться по сети.

На рис. 6 представлена структура построенной нейросетевой модели.

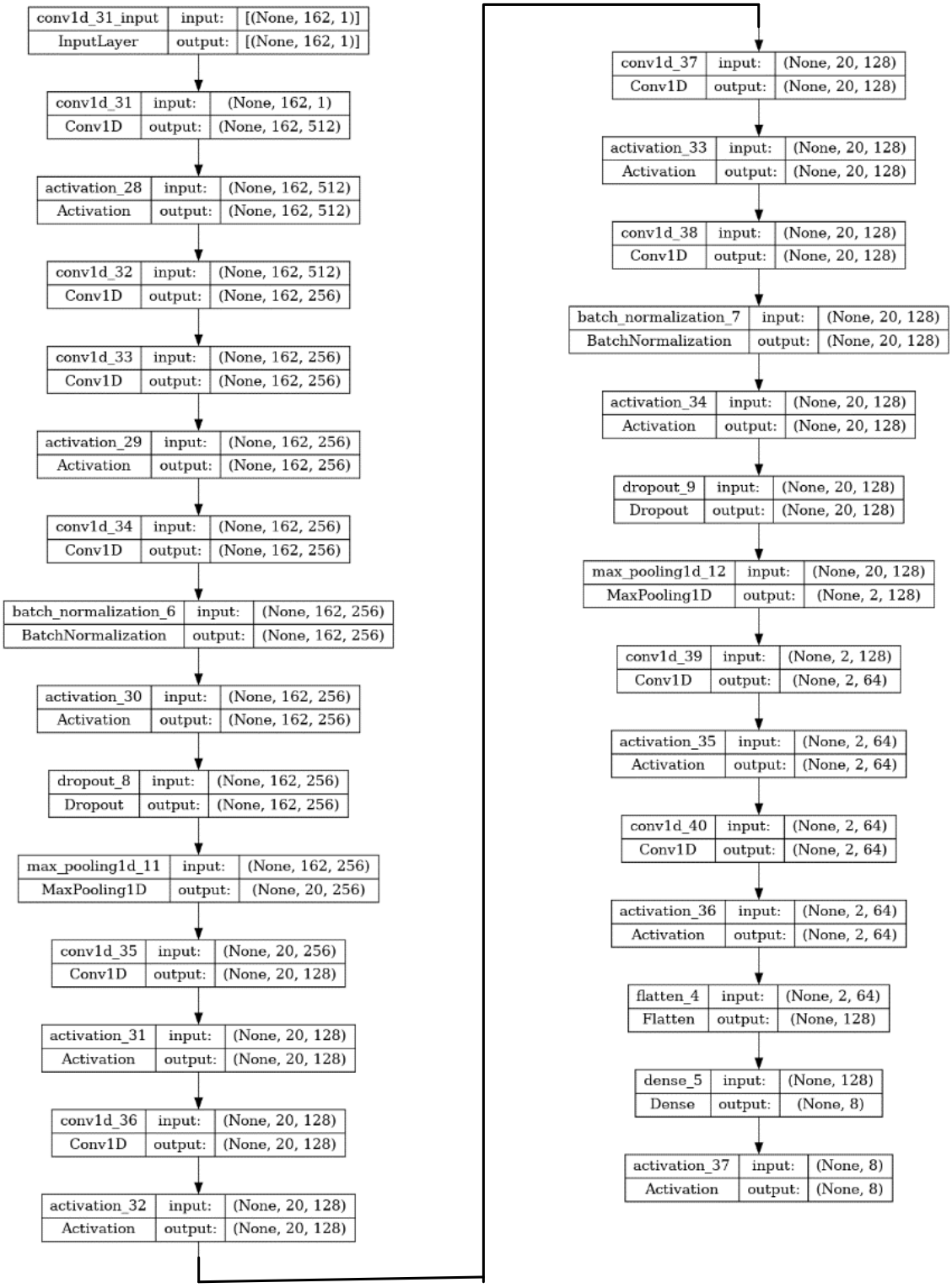


Рис. 6. Структура построенной нейросетевой модели

### Экспериментальные исследования построенной модели

На начальном этапе исследований произведена оценка точности, полноты и F-меры построенной модели при помощи функции classification_report библиотеки sklearn. Результаты оценки представлены на рис.7.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| angry | 0.92 | 0.87 | 0.89 | 1965 |
| calm | 0.86 | 0.93 | 0.89 | 214 |
| disgust | 0.75 | 0.77 | 0.76 | 1884 |
| fear | 0.82 | 0.79 | 0.80 | 1943 |
| happy | 0.79 | 0.83 | 0.81 | 1832 |
| neutral | 0.76 | 0.83 | 0.79 | 1749 |
| sad | 0.81 | 0.77 | 0.79 | 1912 |
| surprised | 0.95 | 0.91 | 0.93 | 663 |
| accuracy | | | 0.82 | 12162 |
| macro avg | 0.83 | 0.84 | 0.83 | 12162 |
| weighted avg | 0.82 | 0.82 | 0.82 | 12162 |

Рис. 7. Результаты оценки точности, полноты и F-меры построенной модели

Далее была построена матрица классификации для построенной нейросетевой модели распознавания эмоциональной окраски речи.

Таблица 2. Матрица классификации на тестовой выборке

| **Истинные классы (количество записей, точность классификации)** | **Предсказанные классы** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | злость | спокойствие | отвращение | страх | счастье | нейтральная эмоция | печаль | удивление |
| злость (1965; 87,38 %) | 1717 | 0 | 104 | 53 | 57 | 21 | 10 | 3 |
| скука (214; 88,79 %) | 0 | 190 | 1 | 0 | 0 | 11 | 12 | 0 |
| отвращение (1884; 78,98 %) | 59 | 3 | 1488 | 57 | 58 | 73 | 137 | 9 |
| страх (1943; 75,66 %) | 42 | 0 | 93 | 1470 | 74 | 69 | 192 | 3 |
| счастье (1832; 77,29 %) | 82 | 0 | 131 | 83 | 1416 | 79 | 29 | 12 |
| нейтральная эмоция (1749; 80,27 %) | 6 | 7 | 105 | 35 | 26 | 1404 | 166 | 0 |
| печаль (1912; 82,90 %) | 2 | 14 | 101 | 102 | 10 | 96 | 1585 | 2 |
| удивление (663; 91,86 %) | 6 | 0 | 12 | 9 | 11 | 7 | 9 | 609 |

Как видно из представленной таблицы, наиболее точно построенная модель определяет эмоции «удивление», «скука», «злость» и «печаль».

Таким образом, в результате проведенных исследований удалось достигнуть точности нейросетевой модели в 81,64 %. На ее точность оказывали влияние выбранные значения гиперпараметров.

## Заключение

Результаты проведенных исследований показали, что построенная нейросетевая модель является эффективной с точки зрения цели моделирования – с высокой точностью определять эмоциональную окраску речи человека. Такая модель может быть эффективно использована в составе интеллектуальной системы поддержки принятия решений. Внедрение подобной системы в различные организации позволит снизить нагрузку на сотрудников, обрабатывающих входящие звонки (например, колл-центры, службы поддержки и пр.). В перспективе с целью развития этого направления планируется построение других интеллектуальных моделей, в частности новых нейросетевых моделей, сравнение полученных результатов с предыдущими, а также создание прототипа программного комплекса, позволяющего в автоматическом режиме распознавать эмоциональную окраску речи человека.

## Список литературы

1. Исследование каналов обращений в службу поддержки клиентов // Helpdesk-система учета заявок OKDESK. URL: https://okdesk.ru/blog/research-itoutsourcing (дата обращения: 12.12.2023).

2. Романюк А.Г. Использование глубокого обучения нейросети для распознавания голосовых команд пользователя / А.Г. Романюк, А.Н. Смирнов, В.М. Антонова // Журнал радиоэлектроники [электронный журнал]. - 2019. - № 11. Режим доступа: http://jre.cplire.ru/jre/nov19/18/text.pdf. DOI 10.30898/1684-1719.2019.11.18 (дата обращения 12.12.2023).

3. Савченко Л. В. Система постановки произношения на основе сверточных нейронных сетей и информационной теории восприятия речи / Л. В. Савченко // Информационные технологии. – 2019. – Т. 25, № 5. – С. 313-319.

4. Вашкевич М.И., Рушкевич Ю.Н. Обзор систем автоматического детектирования речевых нарушений у пациентов с боковым амиотрофическим склерозом (БАС) // Речевые технологии/Speech Technologies. - 2020. - №1-2. URL: https://cyberleninka.ru/article/n/obzor-sistem-avtomaticheskogo-detektirovaniya-rechevyh-narusheniy-u-patsientov-s-bokovym-amiotroficheskim-sklerozom-bas (дата обращения: 14.12.2023).

5. Чучупал В. Я. Неявная модель произношения для автоматического распознавания речи // Речевые технологии/Speech Technologies. 2018. №1-2. URL: https://cyberleninka.ru/article/n/neyavnaya-model-proiznosheniya-dlya-avtomaticheskogo-raspoznavaniya-rechi (дата обращения: 14.12.2023).

6. Scherer K.R. What are emotions? And how can they be measured? // Social science information. – 2005. – Т. 44. – №. 4. – P. 695-729.

7. Галунов В.И. О возможности определения эмоционального состояния по речи // Речевые технологии. – 2008. – № 1. – С. 60–66.

8. Cornelius R.R. The science of emotion: Research and tradition in the psychology of emotions. – Prentice-Hall, Inc, 1996.

9. Calvo R.A. et al. (ed.). The Oxford handbook of affective computing. – Oxford Library of Psychology, 2015.

10. Чернова, А.М. Мягкие вычисления: понятие, классификация, решаемые задачи / А.М. Чернова // Международная конференция по мягким вычислениям и измерениям. – 2015. – Т. 2. – С. 201-205.

11. Абдулаева, З.И. Применение нечетких множеств и мягких вычислений в медицинской статистике / З.И. Абдулаева // NovaInfo.Ru. – 2016. – Т. 1. – № 51. – С. 236-240.

12. Катасёв, А.С. Гибридная нейронечеткая модель интеллектуального анализа данных для формирования баз знаний мягких экспертных диагностических систем / А.С. Катасёв, Ч.Ф. Ахатова // Наука и образование: научное издание МГТУ им. Н.Э. Баумана. – 2012. – № 12. – С. 34.

13. Глова, В.И. Мягкие вычисления / В.И. Глова, И.В. Аникин, А.С. Катасёв, М.А. Кривилев, Р.И. Насыров. – Казань, 2010. – 206 с.

14. Катасёв, А.С. Распознавание рукописных символов на базе искусственной нейронной сети / А.С. Катасёв, Д.В. Катасёва, А.П. Кирпичников // Вестник Технологического университета. – 2015. – Т. 18. – № 11. – С. 173-176.

15. Доросинский, Л.Г. Сравнительный анализ классических методов и методов машинного обучения при решении задач классификации радиолокационных изображений / Л.Г. Доросинский, С.С. Иванов // Ural Radio Engineering Journal. – 2022. – Т. 6. – № 3. – С. 310-323.

16. Катасёв, А.С. Разработка нейросетевой системы классификации электронных почтовых сообщений / А.С. Катасёв, Д.В. Катасёва // Вестник Казанского государственного энергетического университета. – 2015. – № 1 (25). – С. 68-78.

17. Salamon, J. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification / J. Salamon, J. Bello // DeepAI : [сайт]. – URL: https://deepai.org/publication/deep-convolutional-neural-networks-and-data-augmentation-for-environmental-sound-classification (дата обращения: 01.11.2022).

18. Karol, J.P. Environmental sound classification with convolutional neural networks / J.P. Karol // 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP): [сайт]. – URL: https://ieeexplore.ieee.org/document/7324337 (дата обращения: 01.11.2022).

19. Семенюк, В.В. Разработка алгоритма распознавания эмоций человека с использованием сверточной нейронной сети на основе аудиоданных / В.В. Семенюк, М.В. Складчиков // Информатика. – 2022. – Т. 19. – № 4. – С. 53-68.

20. Балыкин, А.Ф. Моделирование динамических характеристик аналоговых аудио компрессоров с использованием рекуррентных нейронных сетей / А.Ф. Балыкин // Процессы управления и устойчивость. – 2022. – Т. 9. – № 1. – С. 175-183.

21. Багаев, И.И. Анализ понятий нейронная сеть и сверточная нейронная сеть, обучение сверточной нейросети при помощи модуля Tensorflow / И.И. Багаев // Математическое и программное обеспечение систем в промышленной и социальной сферах. – 2020. – Т. 8. – № 1. – С. 15-22.

22. Катасёв, А.С. Интеллектуальный анализ временных рядов в системах диагностики и поддержки принятия решений / А.С. Катасёв, Д.В. Катасёва // Поиск эффективных решений в процессе создания и реализации научных разработок в российской авиационной и ракетно-космической промышленности. Международная научно-практическая конференция. – 2014. – С. 481-483.

23. Барышев, Д.А. Распознавание эмоций человека по речи с помощью рекуррентной нейронной сети / Д.А. Барышев, А.С. Зубанков // NovaUm.Ru. – 2022. – № 40. – С. 12-14.

24. Тупицин, Г.С. Оценка мягкой маски с использованием рекуррентной нейронной сети для подавления шума в речевых сигналах / Г.С. Тупицин, А.И. Топников // Цифровая обработка сигналов. – 2018. – № 4. – С. 45-49.

25. Петрин, Д.А. Улучшение качества моделей машинного обучения в задачах классификации изображений на основе метода аугментации данных / Д.А. Петрин, С.С. Гришунов, Ю.С. Белов // Известия Института инженерной физики. – 2021. – № 1 (59). – С. 56-60.

26. RAVDESS Emotional speech audio. Emotional speech dataset. – Kaggle: [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio (дата обращения: 01.12.2023).

27. Livingstone SR, Russo FA The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English / SR Livingstone. – Текст : электронный // PLoS ONE : [сайт]. – URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391 (дата обращения: 01.12.2023).

28. Surrey Audio-Visual Expressed Emotion (SAVEE). – Kaggle: [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee (дата обращения: 01.12.2023).

29. Toronto emotional speech set (TESS). – Kaggle: [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess (дата обращения: 01.12.2023).

30. Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D). – Kaggle: [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/ejlok1/cremad (дата обращения: 01.12.2023).

# APPLICATION OF NEURAL NETWORK MODELS FOR SOLUTION TASKS OF SPEECH EMOTION RECOGNITION

***D.I. Karpenkova, A.S. Katasev***

Kazan National Research Technical University named after A. N. Tupolev-KAI
10, st. Karl Marx, Kazan, 420111, Russian Federation

**Annotation.** The paper considers the solution to the problem of speech emotion recognition (SER) based on the construction and research of a neural network model. Typical methods of emotion classification are analyzed. To solve the problem, the expediency of using a categorical model of representing emotions as the most effective is justified. Audio recordings of human speech are the object of research. It is proposed to use a neural network model to analyze the values of audio recording parameters, such as spectral coefficients, spectrograms and chromatograms. Several sets of English-language audio data found on the kaggle platform were used as source data for analysis and neural network modeling. The original dataset identifies seven classes (emotions): happiness, surprise, neutral emotion, anger, sadness, fear, disgust. The total number of audio recordings in the generated archive is 48,648. The initial data was presented in the form of audio recordings of various lengths. To train a neural network model, characteristic features were extracted from audio recordings and augmentation was performed. Based on the initial data, the values of 162 parameters of audio recordings were calculated to obtain a single data table for analysis. The process of preparing data for analysis and modeling is described. The data was divided into training and test sets, as well as the construction and study of a neural network model in the form of a convolutional neural network. To assess the effectiveness of the constructed model, an assessment of the accuracy, completeness and F-measure of the constructed model was made. The research results have shown that the model is quite effective and can be used as part of an intelligent decision support system.

**Keywords:** neural network model, speech emotion, audio data analysis, modeling.

Статья представлена в редакцию 18 декабря 2023 г.