УДК 004.021 EDN: OQLNYG

Автоматическое распознавание эмоций из речевого сигнала на основе оптимизированных свёрточных нейронных сетей

Хабибуллоев Мухаммадсодик Сайфуллоевич

Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, г. Казань, РФ

*E-mail: sodik9911@gmail.com

Аннотация. Распознавание человеческих эмоций с помощью машин - сложная задача. Модели глубокого обучения пытаются автоматизировать этот процесс, заставляя машины проявлять способность к обучению. Однако распознавание человеческих эмоций из речевых сигналов с хорошей точностью все еще остается сложной задачей. С появлением алгоритмов глубокого обучения, эта проблема была решена. В данном исследовании мы изучили сверточные нейронные сети и их применение для распознавания человеческих эмоций. Предложена базовая нейросетевая модель, которая имеет два сверточных слоя, за каждым из которых следовал слой объединения, и один полностью связанный слой. Приведены результаты анализа и экспериментов на двух наборах данных: RAVDESS и TESS. Для улучшения результатов и производительности модели были применены разные техники и получены новые архитектуры.

Ключевые слова: Распознавание эмоций, глубокое обучение, нейронные сети, сверточные нейронные сети, LibROSA

Automatic speech emotion recognition based on optimized convolutional neural networks

Khabibulloev Mukhammdsodik Sayfulloevich

Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan

*E-mail: sodik9911@gmail.com

Abstract. Recognising human emotions by machines is a complex task. Deep learning models attempt to automate this process by forcing machines to show learning ability. However, recognising human emotions from speech signals with good accuracy is still challenging. With the advent of deep learning algorithms, this problem has been solved. In this study, we investigated convolutional neural networks and their application to human emotion recognition. A basic neural network model is proposed which has two convolutional layers, each followed by a union layer, and one fully connected layer. The results of the analysis and experiments on two datasets, RAVDESS and TESS, are presented. Different techniques have been applied to improve the results and the performance of the model and new architectures have been obtained.

Keywords: Emotion recognition, deep learning, neural networks, convolutional neural networks (CNN), LibROSA

1. Введение

За последние несколько десятилетий было проведено множество исследований по изучению человеческого мозга и созданию систем, имитирующих человеческий интеллект. Человеческий мозг — сложный орган, который всегда был источником вдохновения для исследований в области искусственного интеллекта (далее ИИ). Нейронные сети человеческого мозга способны усваивать абстрактные концепции высокого уровня на основе информации низкого уровня, обрабатываемой сенсорной периферией. Изучение языка, понимание речи и распознавание лиц — вот некоторые примеры, демонстрирующие замечательную силу человеческого мозга в изучении понятий высокого уровня. Основная цель ИИ — разработка интеллектуальных систем, способных генерировать рациональные мысли и поведение, аналогичные человеческому мышлению и действиям.

Человеческий мозг использует всю лингвистическую и паралингвистическую информацию, чтобы понять основной смысл высказываний и эффективно общаться. На самом деле любой дефицит в восприятии паралингвистических особенностей отрицательно сказывается на качестве общения. Это подчеркивает важность распознавания эмоциональных состояний речи в эффективном общении. Следовательно, разработка машин, которые понимают паралингвистическую информацию, такую как эмоции, необходима для установления четкого, эффективного и похожего на человека общения.

Существует множество алгоритмов машинного обучения, которые были исследованы для классификации эмоций на основе их акустических характеристик в речевых высказываниях. Учитывая тот факт, что акустические характеристики эмоций в речевых сигналах различаются у говорящих, полов, языков и культур [1], нет единого мнения об акустических признаках эмоций [2, 3]. Использование моделей глубокого обучения является одним из разумных подходов к решению этой проблемы.

Есть три основных этапов для распознавания эмоций из речевого сигнала: предварительная обработка аудиосигнала, выделение и извлечение конкретных спектральных признаков и наконец классификация. Нулевая скорость перехода, спектральный поток, хроматограмма, MFCC (частотные Мел-кепструальные коэффициенты), полиномиальные признаки и т. д., являются одними из наиболее известных признаков, созданных вручную для классификации звука.

В данной работе мы предлагаем модель для распознавания эмоций из речевых сигналов на основе сверточных нейронных сетей. В качестве признака, извлекаемого из аудиосигнала мы выбрали MFCC. Обучение и тестирование модели были проведены с использованием двух открытых наборах данных: RAVDESS[4] и TESS[5].

2. Предложенная модель

В последние годы задача распознавание речи стало популярным среди исследователей в области анализа данных и искусственный интеллект. Существуют различные модели машинного и глубокого обучения для решения задач связанные с распознаванием эмоций из речевого сигнала. Однако наличие настраиваемых гиперпараметров в моделях машинного и глубокого обучения стала проблемой при исследовании, так как при не правильном настраивании этих параметров эффективность модели может снизиться. В связи с этим в данной работе мы предложили оптимизированную модель на основе сверхточных нейронных сетей для распознавания эмоций из речевого сигнала.

На рисунке 1 представлены основные этапы предложенной методики. Он состоит из различных шагов: сбор данных, предварительная обработка аудио-сигналов, разделение набора данных на обучаемых и тестовых, извлечение признаков, обучение и оптимизации модели и распознавания эмоций.

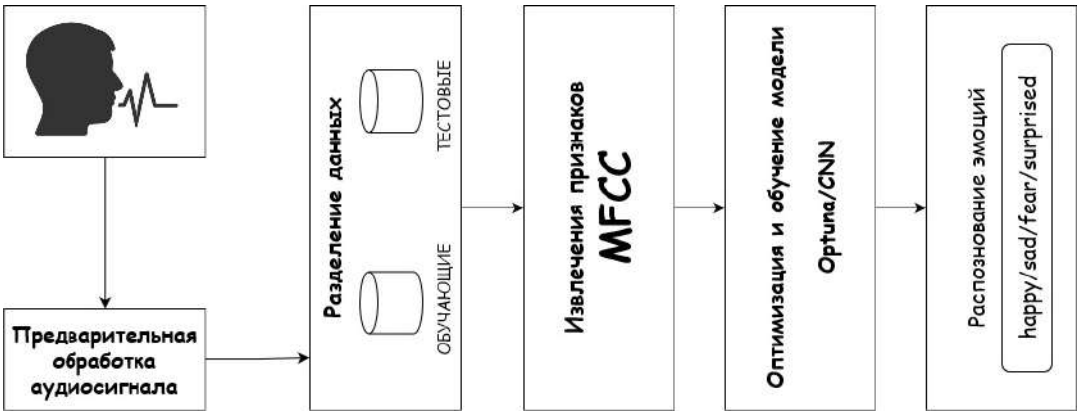

Рисунок 1. Схематическое представление предложенного метода.

2.1. Набор данных

2.1.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

База данных RAVDESS состоит из 24 профессиональных актеров, 12 женщин и 12 мужчин, каждый из которых произносит два лексически связанных предложения.

Речевые эмоции включают выражения «calm», «happy», «angry», «sad», «fearful», «surprise», and «disgust». Для целей нашего эксперимента мы выбрали эмоцию “calm” в качестве “neutral” эмоции. Набор данных включает в себя аудио, аудиовизуальные и видеофрагменты; для нашего исследования мы использовали только аудиофрагменты речи. Этот набор данных очень богат вариациями и содержит североамериканский английский акцент, который присутствует в очень немногих наборах данных.

2.1.2 Toronto Emotional Speech Set (TESS)

Две женщины, носительницы английского языка, выбрали 200 слов для описания каждой из семи эмоций в этом наборе данных: “anger”, “disgust”, “pleasant surprise”, “neutral”, “happiness”, “fear”, и “sorrow”. Две актрисы (26 и 64 лет) произносили список целевых слов, используя фразу-носитель "Скажи слово". Сгенерированный набор данных содержал в общей сложности 2800 образцов. В качестве эмоции “pleasant surprise”, для экспериментального исследования была выбрана эмоция “surprise”. Этот набор данных содержит только записи женского пола и имеет очень хорошее качество звука.

2.2. Предварительная обработка

Аудиосигналы представлены в формате .wav, и эти непрерывные звуковые волны оцифрованы. Они преобразуются в одномерный массив цифровых значений путем дискретной выборки. Такие волны являются однонаправленными и могут использоваться для представления определенной частоты или амплитуды в заданные моменты времени. Выбор частоты дискретизации зависит от теоремы Найквиста-Шеннона, которая показывает взаимосвязь между частотой дискретизации аналого-цифрового преобразователя и частотой дискретизации максимальной частоты сигнала. В нем говорится, что для сбора и восстановления всей информации, присутствующей в непрерывном сигнале, частота дискретизации должна быть в два раза больше, чем частота непрерывного сигнала. Таким образом, полная волна может быть записана в массив без каких-либо шумов или затухания только в том случае, если мы следуем теореме Найквиста-Шеннона. Тем не менее, мы выбрали библиотеку *LibROSA*, которая представляет собой пакет обработки звука и музыки на Python с предопределенными функциями. *LibROSA* нормализует данные, чтобы значения, которые могут быть представлены в массиве, находились в диапазоне от -1 до 1.

LibROSA использует частоту дискретизации по умолчанию 22050 Гц, что помогает поддерживать небольшой размер массива за счет снижения качества звука, но значительного сокращения времени обучения. Поэтому, когда мы загружаем звук с помощью функции *librosa.load()* и для параметра моно установлено значение *true*, он объединяет два канала стереофонического аудиосигнала, чтобы создать одномерный массив *numpy*, тем самым создавая монофонический аудиосигнал. Наконец, после загрузки аудио и представления его в виде массива значений из него извлекаются различные типы признаков.

2.3. Извлечение признаков

Обработка сигналов и извлечение признаков является важным шагом для систем распознавания речи. Существует достаточно большое количество методов, для того чтобы с помощью вектора признаков представить речевой сигнал. Например, Linear Prediction Coding *(LPC)*, Mel-Frequency Cestrual Coefficients *(MFCC)* [7]. В данной работе в качестве метода извлечения признаков мы выбрали MFCC. *MFCC* - это один из методов который используется для анализа речи путем извлечения критических данных и признаков из подмножества речевых данных. Признаки MFCC вычисляются с помощью линейно разнесенных частотных фильтров на низких частотах и логарифмически разнесенных частотных фильтров на высоких частотах.

2.4. Архитектура предложенной модели

Наша модель была обучена и оценена с помощью 5-кратной кросс-валидации. То есть, данные были разделены на 5 частей. Первая часть использовалась в качестве тестового набора, в то время как остальные части использовались для обучения наших моделей. Затем вторая часть использовалась для тестирования наших моделей, а остальные части использовались для обучения, и так далее. Для уменьшения перегрузки и негативного эффекта малого размера баз данных, наборы данных были дополнены путем добавления белого гауссовского шума с отношением сигнал/шум (SNR) +15 к каждому аудио сигналу либо 10 раз, либо 20 раз**.**

Базовая архитектура глубокой нейронной сети, которая была реализована в данном исследовании, представляла собой сверточную нейронную сеть с двумя сверточными слоями и одним полностью связанным слоем с 1024 скрытыми нейронами. В зависимости от количества классов, для оценки распределения вероятностей классов использовался либо 5-полосный, либо 7-полосный блок softmax. За каждым сверточным

слоем следовал слой с максимальным или средним объемом. Прямолинейные блоки *(ReLU*) использовались в сверточных и полностью связанных слоях в качестве функций активации для внесения нелинейности в модель. Начальный размер ядра сверточных слоев был установлен на 5x5 с шагом 1. Начальное количество ядер было установлено на 8 и 16 для первого и второго сверточных слоев, соответственно. Размер ядра объединяющих слоев был установлен на 2x2 с шагом 2. В качестве функции потерь использовалась перекрестная энтропия, а для минимизации функции потерь на мини-пакетах обучающих данных использовался оптимизатор Adam. Размер mini batch был установлен на 512. Количество итераций обучения составило 100. Также, мы включили алгоритм отсева (dropout) в полностью подключенный слой (*FCL*), чтобы улучшить производительность наших сетей всякий раз, когда диагностировались симптомы переобучение. На рисунке 2 показаны основные структурные блоки модели CNN реализованной в данной работе.

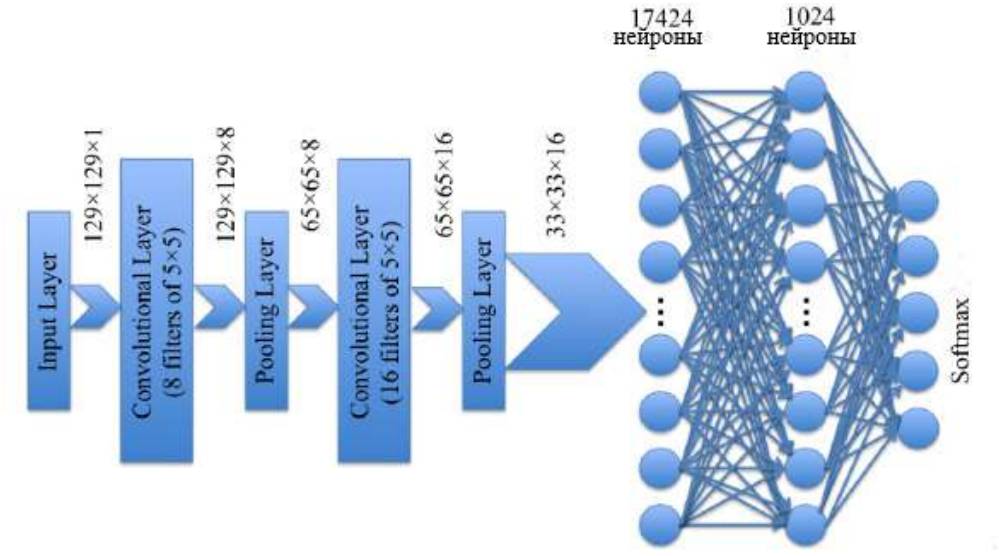

Рисунок 2. Базовая архитектура CNN, используемая в данном исследовании для распознавания речевых высказываний на основе их эмоционального состояния.

Для оптимизации модели используется метод оптимизации *Optuna*. Было адаптировано набор значений для различных параметров *CNN*: количество эпох, значение для dropout, размер mini batch, размер ядра и количество фильтров. В таблице 1 представлены значения параметров, которые были адаптированы для CNN

Таблица 1. Значения параметров для OPCNN.

Параметр	Значения
Dropout (%)	В диапазоне от 0.1 до 0.6
количество фильтра	8, 16
Размер ядра	5, 2
Размер pooling	2
Размер batch	512
Количество эпох	100

Мы использовали *TensorFlow* (библиотеку с открытым исходным кодом, написанную на *Python* и *C++* [6]) в качестве основы программирования для реализации наших CNN-моделей.

4. Вычислительные эксперименты

Мы провели несколько экспериментов на каждой базе данных в зависимости от языка и пола. Мы начали наше исследование с реализации базовой архитектуры CNN, представленной в главе 2. Затем мы изменили гиперпараметры, такие как размер ядер свертки и вероятность удаления алгоритма отсева, что повлияло на производительность моделей.

4.1. Результаты

4.1.1. RAVDESS

Мы провели несколько экспериментов с базой данных *RAVDESS*. Во всех экспериментах наши сети обучали данные с точностью 100%. Однако точность на проверочных данных была различной для разных архитектур. В таблице 2 представлены архитектуры и их соответствующая точность на тестовых данных. Точность на проверочных данных — это средняя точность тестовых данных по 5 сложениям оценки перекрестной валидации. На рисунке 3 показана точность моделей CNN при обучении и проверки для каждой итерации обучения.

Как видно из рисунка, первое значительное увеличение производительности нашей сети произошло при использовании алгоритма отсева. Применение отсева увеличило точность проверки с 57.4% до 77.7%. Увеличение размера окна первого сверточного слоя с 5*5 до 10*10 также улучшило производительность сети. Кроме того, использование среднего объединения вместо максимального объединения повысило производительность.

Таблица 2. Параметры архитектур и результатов экспериментов, проведенных на базе данных RAVDESS; f1 и f2 - размер ядер в первом и во втором сверточном слое.

f1	f2	pooling	p(dropout)	augumentation	epoch	CV accuracy, %
5x5	5x5	Max	0	10x	100	57.4
5x5	5x5	Max	0.5	10x	100	77.7
10x10	5x5	Max	0.5	10x	100	85.29
10x10	5x5	Average	0.5	10x	100	87.58
5x5	5x5	Max	0.5	20x	100	95
10x10	5x5	Max	0.5	20x	100	96.5
10x10	5x5	Average	0.5	20x	100	96.2
10x10	5x5	Average	0.5	20x	800	99.5
10x10	5x5	Average	0.5	20x	4000	99.83

Второй значительный прирост произошел при увеличении числа дополнений данных с 10 до 20 раз. Это увеличение дало точность проверки 95%. Увеличение размера окна первого сверточного слоя повысило производительность с 95% до 96.5%. Использование среднего объединения вместо максимального объединения изменило производительность с 96.5% до 96.2%. В совокупности, самая высокая производительность была связана с архитектурой с 10*10 ядрами в первом сверточном слое, 5*5 ядрами во втором сверточном слое, среднее объединение, отсев с $p = 0.5$ и обучающие данные с 20-кратным увеличением.

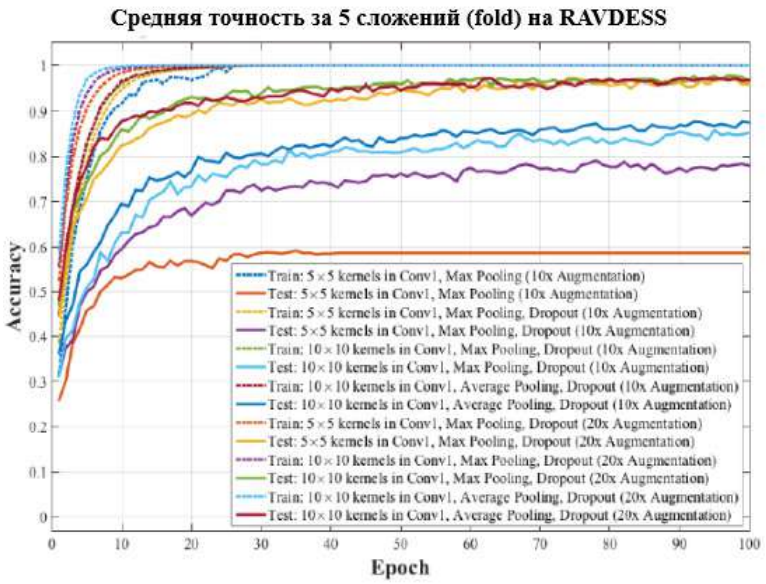

Рисунок 4. Средняя точность CNN на наборе данных RAVDESS.

4.1.2. TESS

В таблице 3 приведены результаты экспериментов, проведенных на базе данных TESS. Результаты показывают, что сеть с 11 ядрами в первом сверточном слое, 5x5 ядрами во втором сверточном слое, средним объединением, отсевом с $p = 0.5$ и 20-кратным увеличением обучающих данных имеет наилучшую производительность на этой базе данных. На рисунке 5 показана средняя точность работы CNN в течение 5 складок перекрестной валидации на данных с 10-кратным увеличением (изображение (а)) и 20-кратным увеличением (изображение (б)) в течение итераций обучения. Как видно, наши сети с различными архитектурами смогли обучиться на обучающих данных. Однако производительность сетей на тестовых данных варьируется в зависимости от различных архитектур. Как и в случае с базой данных RAVDESS, интеграция отсева в архитектуру повысила производительность сети. Кроме того, увеличение размера окна ядра в первом сверточном слое вместе с объединением средних повысило производительность сети, особенно когда обучающие данные были увеличены в 20 раз.

Таблица 3. Параметры архитектур и результатов экспериментов, проведенных на базе данных TESS; f1 и f2 - размер ядер в первом и во втором сверточном слое.

f1	f2	pooling	p(dropout)	augumentation	epoch	CV accuracy, %
5x5	5x5	Max	0	10x	100	48.5
5x5	5x5	Max	0.5	10x	100	61.4
10x10	5x5	Max	0.5	10x	100	70.2
10x10	5x5	Average	0.5	10x	100	71.6
11x11	5x5	Average	0.5	10x	100	73.4
5x5	5x5	Max	0	20x	100	52
5x5	5x5	Max	0.5	20x	100	63.8
10x10	5x5	Max	0.5	20x	100	72.65
10x10	5x5	Average	0.5	20x	100	74.56
11x11	5x5	Average	0.5	20x	100	81.5
11x11	5x5	Average	0.5	20x	800	95.67
11x11	5x5	Average	0.5	20x	4000	98.2

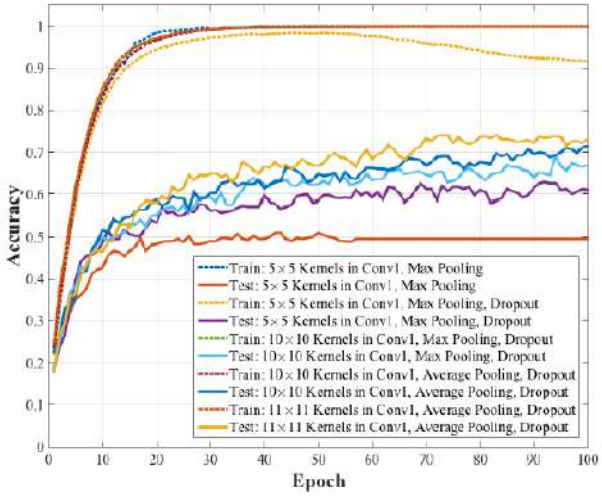

Рисунок 5(а). Средняя точность CNN с 10 кратным увеличением данных на наборе данных TESS.

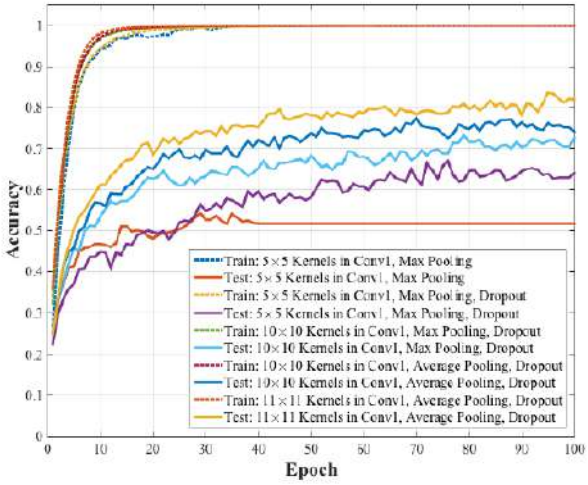

Рисунок 5(б). Средняя точность CNN
20 кратным увеличением данных
наборе данных TESS.

5. Заключение

В данном исследовании мы применили сверточные нейронные сети (*CNN*) для классификации эмоциональных состояний речевых высказываний, используя два общепринятые базы данных [5, 6]. *CNN* имели два сверточных слоя, за каждым из которых следовал слой объединения, и один полностью связанный слой. Для оценки распределения вероятностей классов в выходном слое использовался блок *softmax*. В качестве функции потерь для измерения ошибки оценки использовалась перекрестная энтропия, а для минимизации функции потерь применялся оптимизатор Адама. Все речевые сигналы были преобразованы в широкополосные спектрограммы и поданы в *CNN* в качестве входных данных.

CNN показали хорошие результаты на обучающих данных для всех баз данных. Однако производительность на тестовых данных различалась в разных архитектурах и базах данных.

Результаты данного исследования продемонстрировали компетентность CNN в изучении эмоциональных особенностей речевых сигналов на основе их низкоуровневого представления с использованием широкополосных спектрограмм независимо от пола и языка говорящих. Несмотря на значительный успех наших сетей, все еще существуют возможности для дальнейшего совершенствования. В силу того, что широкополосные спектрограммы не могут разрешить отдельные гармоники, то есть им не хватает некоторой незернистой спектральной информации. Для дальнейшей работы мы предлагаем включить узкополосные спектрограммы в обучающие данные в качестве второго канала входа и сформировать многоканальные обучающие данные. Это может уменьшить количество эпох обучения или количество дополнений данных, которые необходимы для достижения высокой производительности. производительности.

Список литературы

1. Louis Ten Bosch. Emotions, speech and the asr framework / Louis Ten Bosch // Speech Communication. – 2003. – № 40(1-2). – С. 213-225.
2. Moataz El Ayadi. Survey on speech emotion recognition: Features, classification schemes, and databases / Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray // Pattern Recognition. – 2011. – № 44(3). – С. 572-587.
3. Bjorn Schuller. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge / Bjorn Schuller, Anton Batliner, Stefan Steidl, Dino Seppi // Speech Communication. – 2011. – № 53(9-10). – С. 1062-1087.
4. Kaggle.com: сайт. – 2022. – URL: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotionalspeech-audio/code/ (дата обращения 04.07.2022)
5. Kaggle.com: сайт. – 2022. – URL: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speechset-tess/ (дата обращения 04.07.2022)
6. TensorFlow.org: сайт. – 2022. – URL: https://www.tensorflow.org/ (дата обращения 01.07.2022)
7. Аксёнов, О. Д. Метод мел-частотных кепстральных коэффициентов в задаче распознавания речи / Аксёнов О. Д. // Электронные системы и технологии: 55-я юбилейная конференция аспирантов, магистрантов и студентов, Минск, 22-26 апреля 2019 г.: сборник тезисов докладов / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2019. – С. 45-46.