ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРИНЯТИЕ РЕШЕНИЙ
ARTIFICIAL INTELLIGENCE AND DECISION MAKING

УДК 004.89 **DOI: 10.18413/2518-1092-2022-7-3-0-5**

Демин О.Д.
Лаптев А.А.

МОДЕЛЬ ОПРЕДЕЛЕНИЯ ПСИХОЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ЧЕЛОВЕКА НА ОСНОВЕ АУДИО И ВИДЕО ДАННЫХ

Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Кронверкский пр., д. 49, г. Санкт-Петербург, 197101, Россия

e-mail: olegkastrategka@gmail.com, nickname.avast@gmail.com

Аннотация
Определение психоэмоционального состояния человека находит применение в различном количестве задач от диагностирования заболеваний до предупреждения аварийных дорожных ситуаций. Учитывая важность и разнообразие задач, где применяется определение психоэмоционального состояния человека, существует множество методов и подходов, но большинство из них использует только один невербальный признак, например голос или выражения лица человека. При этом одновременное использование нескольких невербальных признаков может увеличить точность по сравнению с методами, использующие только один невербальный признак. В данной работе разработан метод, использующий голос и выражение лица в качестве невербальных признаков. На основе литературного обзора и анализа существующих методов были выбраны лучшие методы и подходы, использующие голос или выражение лица, проведено сравнение разрабатываемого метода с существующими решениями. Разработанный метод позволил улучшить точность CMCNN, но при этом его Accuracy было меньше в сравнении с DisVoice + SVM. Для увеличения точности разработанного метода было предложено использовать средневзвешенное или использовать более сложную модель, способную выявить взаимосвязи между голос и выражением лица человека.
Ключевые слова: психоэмоциональное состояние; анализ эмоций; невербальные признаки; методы определения психоэмоционального состояния
Для цитирования: Демин О.Д., Лаптев А.А. Модель определения психоэмоционального состояния человека на основе аудио и видео данных // Научный результат. Информационные технологии. – Т.7, №3, 2022. С. 43-56. DOI: 10.18413/2518-1092-2022-7-3-0-5

Demin O.D.
Laptev A.A.

MODEL FOR PERSON PSYCHOEMOTIONAL STATE DETERMINATION USING AUDIO AND VIDEO DATA

Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, 49 Kronverkskiy prospekt, St. Petersburg, 197101, Russia

e-mail: olegkastrategka@gmail.com, nickname.avast@gmail.com

Abstract
Determination of the person’s psychoemotional state finds application in a variety of tasks from the diagnosis of diseases to the prevention of emergency traffic situations. Given the importance and variety of tasks where it used, there are many methods and approaches of it exists, but most of them use only one nonverbal feature, such as a person's voice or facial expressions. At the same time, the simultaneous use of several nonverbal signs can increase accuracy compared to methods using only one nonverbal sign. In this paper, such method has been developed that uses voice and facial expression as nonverbal signs. Based on the literature review and analysis of

existing methods, the best methods and approaches using voice or facial expression were selected, the developed method was compared with existing solutions. The developed method made it possible to improve the accuracy of CMCNN, but at the same time, in each test, the Accuracy of its own method was less in comparison with DisVoice + SVM. For improving developed method weighted average was proposed to use instead of usual average or using more complex model, which is able to detect interrelation between person's voice and facial expression.
Keywords: psychoemotional state; emotion analysis; nonverbal; methods for determining emotional state
For citation: Demin O.D., Laptev A.A. Model for person psychoemotional state determination using audio and video data // Research result. Information technologies. – T.7, №3, 2022. – P. 43-56. DOI: 10.18413/2518-1092-2022-7-3-0-5

ВВЕДЕНИЕ

Психоэмоциональное состояние (ПЭС) человека играет важную роль на протяжении всей его жизни. Оно определяет его поведение в различных жизненных ситуациях, а также способность справиться с поставленной задачей.

Под действиями источников стресса, отрицательно влияющих на ПЭС, человек начинает испытывать различные негативные эмоции, такие как грусть или злость, а также возникают отрицательные физиологические симптомы, например, усталость, снижение продуктивности и концентрации. В некоторых случаях для преодоления негативного состояния, принимают различные алкогольные или наркотические вещества [21].

За 2021 год, исключая декабрь, в Российской Федерации произошло 120670 дорожно-транспортных происшествий, среди которых 10275 (≈ 8,5%) инцидентов, где водитель находился в состоянии алкогольного или наркотического опьянения, что привело к гибели 2753 человека [1].

Влияние источников стресса и негативного состояния могут стать причиной развития психических расстройств, например депрессии [21]. Депрессия сильно влияет на жизнь человека, уменьшая его желание делать что-либо, а также может привести к суициду [19]. Так Всемирная Организация Здравоохранения (ВОЗ) оценило количество взрослых (20+ лет), страдающих депрессией, что составило 5% (больше 229 млн. людей) на 2019 год [26]. В России данная цифра составляет 4,73% (больше 4,5 млн. людей) за тот же год [7].

ПЭС человека проявляется в его поведении, общении, при том, большая часть выражается именно невербальными признаками, такими как, лицевое выражение, тон речи, поза [12]. Именно поэтому разработка новых методов автоматического определения психоэмоционального состояния человека на основе невербальных признаков является актуальной задачей. Большинство рассмотренных методов используют один невербальный признак: голос, выражение лица и др. Но использование нескольких невербальных признаков человека, в частности голос и выражение лица, возможно содержит больше информации о его ПЭС, чем при использовании одного.

ПЭС человека отвечает на вопрос, что человек чувствует или испытывает в определённый промежуток времени. Тогда определение ПЭС человека — это нахождение некоторого «состояния» в зависимости от модели представления психоэмоционального состояния или способ описания состояния человека. Используемая модель представления ПЭС зависит как от предложенного подхода, так и от используемого набора данных, что в свою очередь зависит от поставленной цели.

Так в статье [2] авторы занялись вопросом автоматического определения уровня стресса человека по выражению лица, зафиксированного в видеопотоке. Для решения поставленной задачи они предложили использовать подход под названием Segmentation based Fractal Texture Analysis (SFTA) для извлечения параметров из изображения лица пользователя, которая по заявлению автора не зависит от определения лицевых точек, так как их не использует и при этом не трудо- и время затратна, как другие рельефные методы. Предложенная модель определения эмоций показывает 98,5% по метрике Accuracy на не указанном наборе данных.

Кроме определения уровня стресса человека по выражению лица, зафиксированного в видеопотоке, также можно определять уровень депрессии, чем и занялись авторы в статье [9]. Проблемой существующих методов они назвали нечувствительность к микросокращением лицевых мышц, что вредит точности определения эмоции на основе лицевого выражения. Вследствие этого авторы предложили использовать метод Median Robust Local Binary Patterns from Three Orthogonal Planes (MRLBP-TOP), способный не только определять микросокращения, но и пространственно-временные изменения выражения лица, а также стохастическая модель Dirichlet Process Fisher Vector (DPFV) для получения глобальных характеристик. Модель была обучена и проверена на наборах данных AVEC2013 и AVEC2014 (Continuous Audio/Visual Emotion and Depression Recognition Challenge), где получила точность 7,55 Mean Absolute Error (MAE) и 7,21 MAE соответственно.

Депрессия не единственное заболевание, которое можно диагностировать, используя определение ПЭС человека по видеопотоку, содержащему выражения лица человека. Fei Z. и другие [4] использовали определение ПЭС человека для построения моделей определения эмоций возможной в применении для диагностирования ранних признаков когнитивного заболевания по мимике лица, зафиксированного в видеопотоке.

Также определение ПЭС человека можно использовать для определения усталости водителя по выражению лица водителя, зафиксированного во входном видеопотоке, что и стало целью авторов в статье [25]. В ней авторами был предложен подход в определении усталости водителя за рулём, чтобы уменьшить количество дорожно-транспортных происшествий.

Многие методы определения ПЭС человека не рассчитаны на использование в естественных условиях из-за возникающих сложностей, например, различные условия освещенности. Кроме того, многие наборы данных не сбалансированы по количеству примеров для некоторых эмоций, например, удивление, из-за чего уменьшается точность определения. Решением данной проблемы занялись авторы в статье [28]. Для этого они предложили нейронную модель Convolutional Relation Network для определения эмоций по выражению лица, содержащегося в видеопотоке, на основе few-shot learning. Модель состоит из четырех свёрточных блоков, извлекающие параметры из входных данных, которые затем переводятся в одно параметрическое пространство, описывающее классы эмоций. На основе этого пространства определяется конкретная эмоция из набора возможных: счастье, удивление, грусть и др. Предложенная модель продемонстрировала 56,25% Accuracy на наборе данных RAF-DB, 67,32% Accuracy на наборе данных FER2013 и 54,87% Accuracy на наборе данных SFEW.

Большинство приведённых статей ранее использовали выражение лица человека изображении или на видео, но кроме него можно также использовать голос, что и сделали авторы в статье [10]. В ней авторы указали на существующие проблемы нахождения минимума в частоте голоса, закодированного спектрограммой Мэла, у нейронной сети BP. Для решения этой проблемы они предложили использовать Particle Swarm Optimization Algorithm (PSO). Результатом определения данного подхода становятся эмоция из возможного набора: злость, счастье, грусть, страх или нейтральная эмоция. Данный подход позволил увеличить общую точность определения эмоции по голосу человека с Accuracy в 90,39% до 97,62% на наборе данных CASIA-CESC (Chinese Academic of Science Institute of Automation-Chinese Emotional Speech Corpus).

Большинство методов определения ПЭС человека по голосу используют методы и модели глубокого обучения, для обучения которых требуется время и вычислительные мощности. Поэтому авторы в статье [3] предложили подход определения эмоций по голосу, зафиксированного в аудиопотоке, с помощью метода интерполяции Cubic Spline Interpolation (CSI). Данный подход продемонстрировал 69,07% Accuracy на наборе данных RAVDESS, 92,52% Accuracy на наборе данных Emo-DB и 89,1% Accuracy на наборе данных SAVEE (Surrey Audio-Visual Expressed Emotion).

В изучаемых статьях модели представления ПЭС были как множество конкретных эмоций или состояний (радость, усталость) [2-6, 8-11, 13-17, 20, 23-25, 27-28], конкретное значение какой-то характеристики ПЭС, например, уровень депрессии [9], стресса [13], наличие ранних признаков

когнитивных заболеваний [2, 4], уровень удовлетворённости [15], так и несколько числовых характеристик одновременно, например, активность-валентность-контроль (АВК) [6].

Модели представления можно разделить на 2 группы: дискретные (конкретное состояние из множества возможных) или пространственные (одна или несколько числовых характеристик, описывающих состояние человека). Различные модели представления ПЭС, используемые в статьях, продемонстрированы на рисунке 1.

Можно увидеть, что большинстве работ используется конкретный набор состояний, например, злость, грусть, усталость и др. Меньше всего используют одиночные характеристики такие как уровень стресса, депрессии, удовлетворённости. Несколько характеристик, например АВК, также мало используется для представления ПЭС.

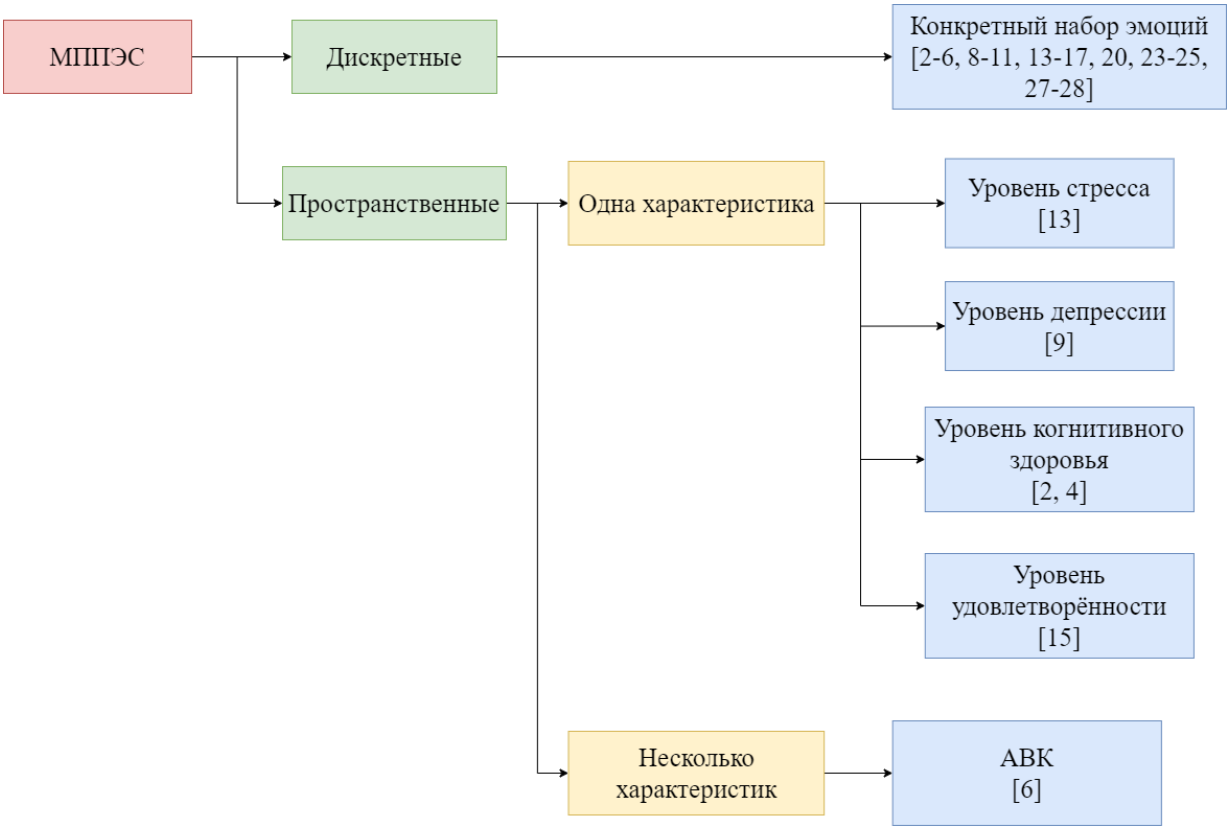

Рис. 1. Модели представления ПЭС, используемые в статьях
Fig. 1. Psychoemotional state representation models

Все описанные работы выше используют невербальный признаки, зафиксированные либо видео-, либо аудиопотоком. Используемые невербальные признаки вместе с их источником представлены на рисунке 2.

Можно увидеть, что чаще всего используется выражение лица человека, реже невербальные признаки, связанные с голосом, например, тон. Самый редкий невербальный признак – походка человека.

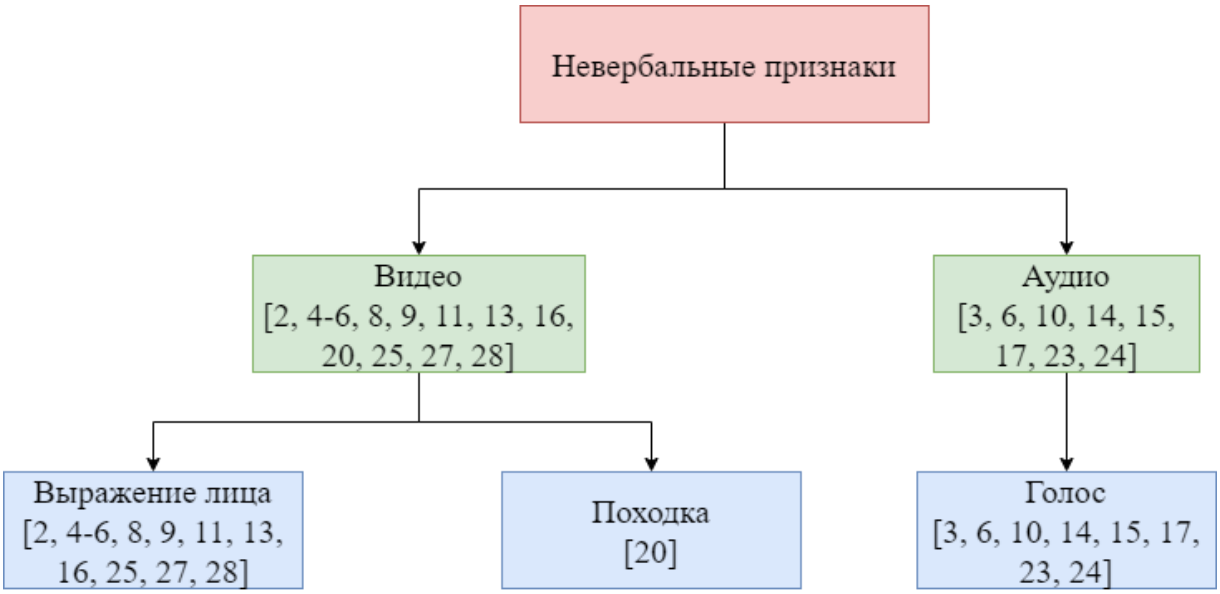

Рис. 2. Используемые невербальные признаки
Fig. 2. Used nonverbal characteristics

Некоторые выше разобранные статьи использовали определение ПЭС как инструмент достижения некоторой поставленной цели (рисунок 3.).

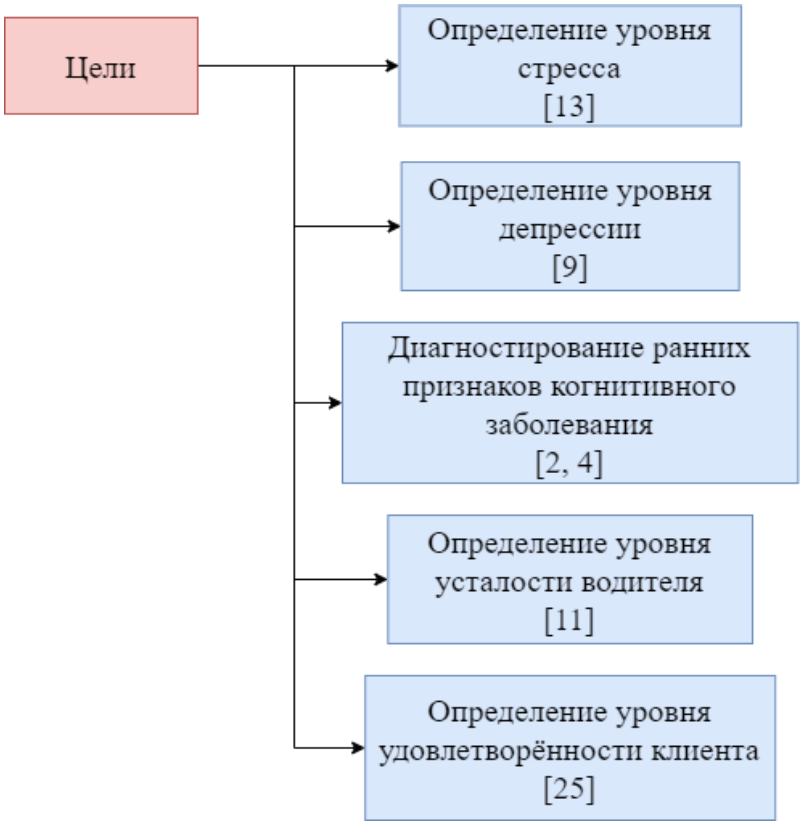

Рис. 3. Цели применения ОПЭСЧ
Fig. 3. Goals of using psychoemotional state determination

В таблице 1 приведены модели и подходы по способности работы в естественных условиях. В ней знак «-» означает, что данный метод или подход не обладает или не проверялся на наличие данного достоинством, а знак «+» наоборот. Пропуск значит, что данный метод или подход не оценивается по данному достоинству.

Таблица 1

Модели и подходы по способности ОПЭСЧ в естественных условиях

Table 1

Models and approaches by ability to be used in natural environment

Модель/Подход	Работает в различных условиях освещенности	Определяет ПЭС при частично закрытом лице	Справляется с фоновым шумом
SFTA [13]	-	-	
MRLBP-TOP+DPFV [9]	+	-	
AlexNet+ЛРД [4]	-	-	
MobileNetV2+SVM [2]	-	-	
CNN+CNN [25]	-	-	
Xception [5]	-	-	
DisVoice+SVM [15]			-
CMCNN [27]	-	-	
CRN [28]	+	-	
RAN [11]	-	+	
ECNN+ЛРД [16]	**+**	**+**	
[8]	-	+	
SOA+PSO [10]			-
TFNN [14]			-
CapsNet [24]			-
Comp-CapsNet+DC-LSTM [17]			-
CSI [3]			-
VFL [20]	**+**		
SFF-NEC+SVM [23]			-
TCN+R-GCN [6]	-	-	-

Другие достоинства моделей и подходов, которые можно выделить у них: требовательность к вычислительным мощностям и к большему количеству обучающих данных. Модели и подходы, рассмотренные с точки зрения этих достоинств приведены в Таблице 2.

Таблица 2

Модели и подходы по требовательности

Table 2

Models and approaches by computing power and data requirement

Модель/Подход	Требовательна (ен) к вычислительным мощностям	Требует большое количество данных для обучения
SFTA [13]	**-**	**-**
MRLBP-TOP+DPFV [9]	-	+
AlexNet+ЛРД [4]	+	+
MobileNetV2+SVM [2]	+	+
CNN+CNN [25]	+	+
Xception [5]	+	+
DisVoice+SVM [15]	**-**	**-**
CMCNN [27]	+	+
CRN [28]	+	-
RAN [11]	+	+
ECNN+ЛРД [16]	+	+

Модель/Подход	Требовательна (ен) к вычислительным мощностям	Требует большое количество данных для обучения
[8]	+	+
SOA+PSO [10]	-	+
TFNN [14]	+	+
CapsNet [24]	+	+
Comp-CapsNet+DC-LSTM [17]	-	+
CSI [3]	-	+
VFL [20]	+	+
SFF-NEC+SVM [23]	-	+
TCN+R-GCN [6]	+	+

Из данных таблиц можно увидеть, что наиболее приспособленными моделями и подходами в работе в естественных условий являются: ECNN+ЛРД [16] и VFL [20]. Большинство моделей и подходов показывают высокую точность определения только в контролируемых условиях.

Также можно заметить, что наименее требовательными моделями и подходами к вычислительным мощностям и количеству данных для обучения являются: SFTA [13] и DisVoice+SVM [15]. Остальные по большей части требуют, как минимум один из показателей.

Одним из важных характеристик моделей и подходов является точность их работы, представленные разными метриками на различных наборах данных. Сравнив их между собой, можно определить лучшие в своих категориях.

В таблице 3 приведены модели и подходы, использующие выражения лица в качестве невербального признака. Прочерком указано отсутствие данных по этому набору данных для модели или подхода. Значения в таблице представлены метрикой Accuracy. Лучшие значения на наборе данных выделены полужирным шрифтом.

Таблица 3

Сравнение точности моделей и подходов, использующих выражения лица

Table 3

Accuracy comparison of models and approaches, which uses facial expression as nonverbal characteristic

Модель/подход	KDEF	CK+	FER2013	RAF-DB	SFEW
AlexNet+ЛРД [4]	88,4%	-	-	-	-
MobileNetV2+SVM [2]	**88,7%**	-	-	-	-
CNN+CNN [25]	-	**97%**	-	-	-
Xception [5]	-	-	64,71%	-	-
CMCNN [27]	-	96,02%	-	77,03%	34,95%
CRN [28]	-	-	**67,32%**	56,25%	**54,87%**
ECNN+ЛРД [16]	-	86,5%	-	**86,2%**	-
[8]	-	95,71%	-	76,95%	-

Как можно заметить на наборе данных KDEF лучше всего себя показала MobileNetV2+SVM [2], на наборе данных CK+ самую высокую точность продемонстрировала CNN+CNN [25], на наборе данных FER2013 лучшую точность продемонстрировала CRN [28], на наборе данных RAF-DB самый высокий результат показала ECNN+ЛРД [16], а на наборе данных SFEW – снова CRN [28]. Больше всего лучших результатов продемонстрировала CRN [28]. Лучшую среднюю Accuracy продемонстрировала CMCNN (41,6%).

Также сравним модели и подходы, использующие голос в качестве невербального признака. Результаты сравнения представлены в таблице 4. Значения представлены в виде метрики Accuracy.

Таблица 4

Сравнение точности моделей и подходов, использующих голос

Table 4

Accuracy comparison of voice-based models and approaches

Модель/подход	IEMOCAP	Emo-DB	RAVDESS
DisVoice+SVM [15]	**67,5%**	91,4%	**83,5%**
TFNN [14]	55,56%	85,15%	-
CapsNet [24]	-	**99,76%**	-
Comp-CapsNet+DC-LSTM [17]	-	-	82,1%
CSI [3]	-	92,52%	69,07%

Как можно заметить из таблицы 4 на наборе данных IEMOCAP лучше всего себя показал DisVoice+SVM [15], на наборе данных Emo-DB самую высокую точность продемонстрировала CapsNet [24], а на наборе данных RAVDESS – DisVoice+SVM [15]. Больше всего лучших результатов показала DisVoice+SVM [15]. Лучшую среднюю Accuracy продемонстрировал DisVoice + SVM (80,8%).

Таким образом, в ходе анализа статей было установлено, что:

1. Большинство моделей и подходов основаны на глубоком обучении;

2. Существует мало работ, посвященных другим невербальным признакам, таким как, походке, позе, жестам рук и другим;

4. Большая часть существующих моделей и подходов для определения ПЭС человека не приспособлены к работе в естественных условиях;

5. Существует мало исследований, в которых для используют одновременно несколько невербальных признаков.

Также по результатам сравнения точности работы различных методов и подходов можно заметить, что лучше всего себя проявили CMCNN и DisVoice + SVM с лучшими средними Accuracy по различным набора данных. Далее они будут использованы для сравнения с разрабатываемым методом.

РАЗРАБАТЫВАЕМЫЙ МЕТОД

В ходе проведения исследования, был разработан следующий метод - метамодель на основе блендинга результатов работы нейронной модели CMCNN и метода DisVoice + SVM.

Функцию гипотезы разработанного метода можно представить следующей формулой (1):

$$a(x) = 0.5 \times b_1(x) + 0.5 \times b_2(x), \quad (1)$$

где $a(x)$ – функция гипотезы собственного метода, $\mathbb{R}^8$;

$b_1(x)$ – функция гипотезы CMCNN, $\mathbb{R}^8$;

$b_2(x)$ – функция гипотезы DisVoice + SVM, $\mathbb{R}^8$.

То есть результатом работы разработанного метода является среднее арифметическое результатов работы CMCNN и DisVoice + SVM.

ИСПОЛЬЗУЕМЫЙ НАБОР ДАННЫХ

Для обучения и оценки данного метода был использован набор данных The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [18].

RAVDESS содержит всего 7356 файлов общим размером 24,8 Гб. Эти файлы содержат вокализированные лексически одинаковые утверждения в нейтральном североамериканском акценте и песни, произведенные 24 профессиональными актерами с различной отыгрываемой эмоцией (нейтральная эмоция, счастье, грусть, злость, страх, удивление, отвращение). Для песен набор эмоцией меньше: нейтральная эмоция, счастье, грусть, злость и страх. При этом каждая отыгранная эмоция варьируется по интенсивности: обычная или сильная.

Пример данных из RAVDESS представлен на Рисунке 4, взятого из [18].

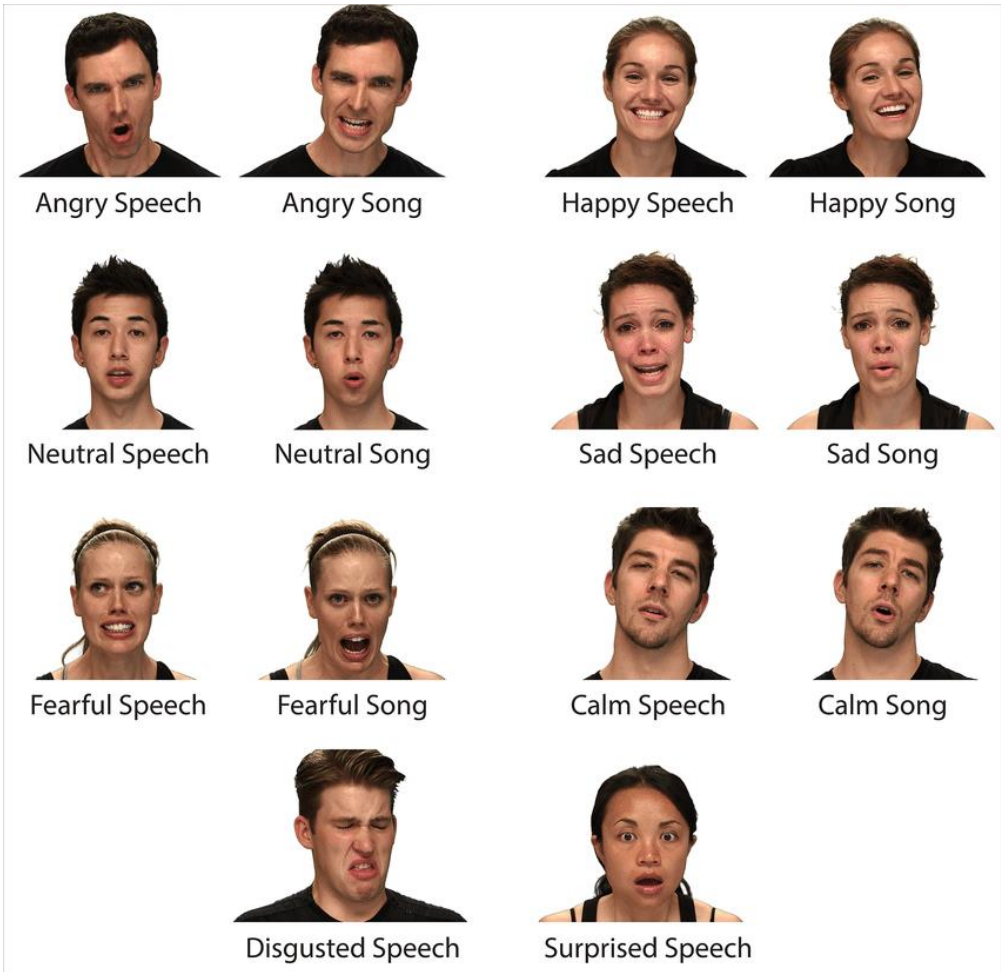

Рис. 4. Пример данных из RAVDESS
Fig. 4. RAVDESS data example

Все данные внутри набора представлены в трёх форматах: только аудио (16 бит, 48 кГц, .wav), аудио и видео (720p H.264, AAC 48 кГц, .mp4), только видео (без звука).

Вся информация о каждой попытке воспроизведения эмоции содержится в названии самого файла в виде числовых идентификаторов, например, “02-01-06-01-02-01-12.mp4”, что означает по порядку появления в строке:

• Модальность (01 – звук и видео, 02 – только видео, 03 – только звук)

• Тип (01 – речь, 02 – песня)

• Эмоция (01 – нейтральная, 02 – спокойствие, 03 – счастье, 04 – грусть, 05 – злость, 06 – страх, 07 – отвращение, 08 – удивление)

• Интенсивность (01 – обычная, 02 – сильная)

• Утверждение (01 – “Kids are talking by the door”, 02 – “Dogs are sitting by the door”)
• Попытка (01 – первая попытка, 02 – вторая)
• Актер (01-24, где нечетное число обозначает мужской пол, а четное женский)

Так как проявление эмоции при исполнении песни и вокализации утверждения может сильно различаться, в данном эксперименте будет использовано подмножество набора данных RAVDESS только с произнесением утверждений, тогда количество используемых файлов при эксперименте 4320 файлов. Набор данных находится в публичном доступе по адресу [18].

Необходимо отметить, что существует риск влияния гипертрофированности эмоций актеров, что может сказаться на точности. Однако данный способ позволяет точно определить какая эмоция отображена человеком, в отличии от естественных ситуаций, где человек может демонстрировать смешанные эмоциональные отклики.

Также данный набор данных прошёл серию оценочных экспериментов точности воспроизведенных эмоций, в которых приняли участие 247 студентов-добровольцев с разным уровнем и направления образования, в том числе в области актерского искусства. Им были предложены видео и аудио фрагменты, для которых они должны были определить, что за эмоция отображена, а также её интенсивность и искренность. В результате была получена средняя точность 73% для собранного набора данных.

СРАВНЕНИЕ РАЗРАБАТЫВАЕМОГО МЕТОДА С ДРУГИМИ МЕТОДАМИ ОПЭСЧ

Для сравнение разрабатываемого метода была предложена следующая схема, представленная на рисунке 5.

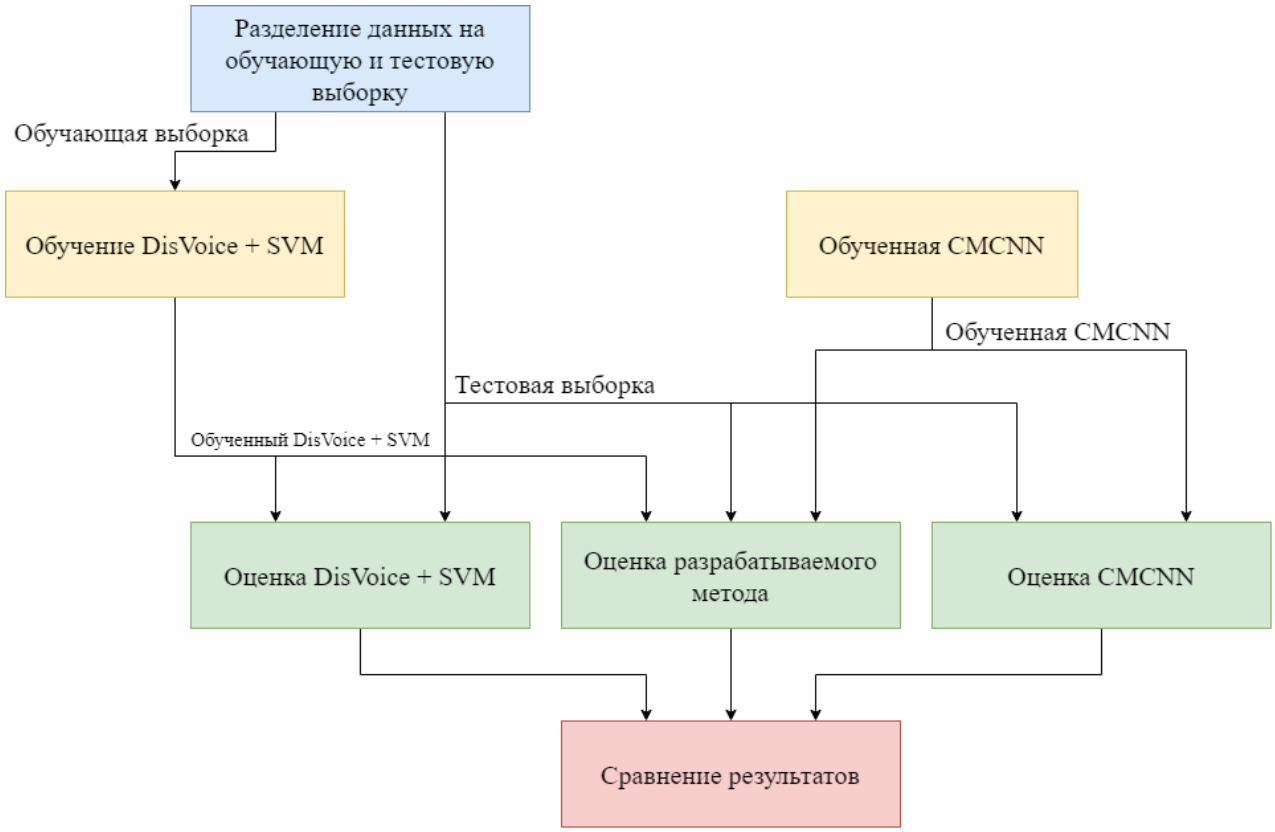

Рис. 5. Общая схема алгоритма проведения эксперимента
Fig. 5. General experiment scheme

В ходе работы было выполнено несколько итераций испытаний для получения средних оценок точности методов.

Результаты сравнения собраны в таблице 5.

Таблица 5

Результаты испытаний

Table 5

Test results

№	CMCNN, Accuracy	DisVoice + SVM, Accuracy	Собственный метод, Accuracy	CMCNN, средняя Accuracy	DisVoice + SVM, средняя Accuracy	Собственный метод, средняя Accuracy
1	62,0%	67,1%	62,5%	61,72%	65,5%	62,02%
2	61,0%	65,2%	61,5%			
3	63,2%	63,7%	63,0%			
4	62,0%	65,5%	61,9%			
5	60,0%	66,0%	61,2%			

Исходя из результатов таблицы можно увидеть, что собственный метод позволил улучшить точность CMCNN в большинстве случаев, но он при этом в каждом испытании Accuracy собственного метода была меньше, чем у DisVoice + SVM. Поэтому средняя Accuracy собственного метода выше, чем у CMCNN, но меньше, чем у DisVoice + SVM.

Разрабатываемый метод считает среднее арифметическое результатов работы нейронной модели CMCNN и метода DisVoice + SVM, поэтому верхний порог результата работы разрабатываемого метода зависит от точности работы менее точного.

В качестве решения данной проблемы можно:

1. Улучшить меньший по точности метод;

2. Использовать среднее взвешенное вместо среднего арифметического, назначая менее точному методу меньший вес.

Также можно разработать более сложную модель на основе глубокого машинного обучения, поскольку среднеарифметическое, так же, как и средневзвешенное, не способно выявить взаимосвязь выражения эмоций по лицу и по голосу.

В качестве более сложной модели можно предложить модели по типу multi-task, как например CMCNN, определяющая взаимосвязи между извлеченными параметрами выражений лица и лицевых точек.

ЗАКЛЮЧЕНИЕ

В данной работе был разработан метод определения ПЭС человека, использующий голос и выражение лица человека в качестве невербальных признаков. В ходе литературного обзора и анализа существующих методов были определены лучшие методы и подходы по голосу или выражению лица человека. На основе сравнения разработанного метода с лучшими методами было установлено, что разработанный метод позволил улучшить точность CMCNN в большинстве случаев, но он при этом в каждом испытании Accuracy собственного метода была меньше, чем у DisVoice + SVM.

Таким образом, для увеличения точности разработанного подхода и получения более достоверных результатов целесообразно:

1. Использовать среднее взвешенное, назначая меньший вес менее точному методу, вместо среднеарифметического;

2. Разработать более сложную модель, способную выявить взаимосвязи между характеристиками изображения лица и голоса.

Список литературы

1. Сведения о показателях состояния безопасности дорожного движения в Российской Федерации. URL: http://stat.gibdd.ru (дата обращения: 21.12.2021).

2. Fei Z., Yang E., Yu L., Li X., Zhou H., Zhou W. A Novel deep neural network-based emotion analysis system for automatic detection of mild cognitive impairment in the elderly // Neurocomputing. 2022. Том 468. С. 306-316.

3. Abdulmohsin H.A., Wahab H.B.A., Hossen A.M.J.A. A novel classification method with cubic spline interpolation // Intelligent Automation and Soft Computing. Том 31. – 2022. - №1. – С. 339-355.

4. Fei Z., Yang E., Li D.D.-U., Butler S., Ijomah W., Li X., Zhou H. Deep convolution network based emotion analysis towards mental health care // Neurocomputing. 2020. Том 388. С. 212-227.

5. Almeida J., Vilaça L., Teixeira I.N., Viana P. Emotion identification in movies through facial expression recognition // Applied Sciences (Switzerland). 2021. Том 11. №15.

6. Tu G., Wen J., Liu H., Chen S., Zheng L., Jiang D. Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models [Formula presented] // Knowledge-Based Systems. 2022. Том 235.

7. GBD Result Tool search result for depression in Russian Federation. URL: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/4ba5e9f86007614b6cc22ce8b7d41f7f (дата обращения: 21.12.2021)

8. Gan C., Xiao J., Wang Z., Zhang Z., Zhu Q. Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention // Image and Vision Computing. Том 117. – 2022.

9. He L., Jiang D., Sahli H. Automatic Depression Analysis Using Dynamic Facial Appearance Descriptor and Dirichlet Process Fisher Encoding // IEEE Transactions on Multimedia. 2019. Том 21. №6. С. 1476-1486.

10. Li J., Qin D. The mutation seagull algorithm optimizes the speech emotion recognition of BP neural network // ACM International Conference Proceeding Series. 2021. С. 160-164.

11. Lyu L., Zhang Y., Chi M.-Y., Yang F., Zhang S.-G., Liu P., Lu W.-G. Spontaneous facial expression database of learners' academic emotions in online learning with hand occlusion // Computers and Electrical Engineering. Том 97. – 2022.

12. Mehrabian A. Nonverbal Communication. New York, 1972. 235 с.

13. Nair P., Subha V., Aneesh R. Non Verbal Behaviour Analysis for Distress Detection Using Texture Analysis // 2018 International CET Conference on Control, Communication, and Computing, IC4 2018. 2018. С. 239-244.

14. Pandey S.K., Shekhawat H.S., Prasanna S.R.M. Attention gated tensor neural network architectures for speech emotion recognition // Biomedical Signal Processing and Control. 2022. Том 71.

15. Parra-Gallego L.F., Orozco-Arroyave J.R. Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments // Digital Signal Processing: A Review Journal. Том 120. – 2022.

16. Prabhu K., SathishKumar S., Sivachitra M., Dineshkumar S., Sathiyabama P. Facial expression recognition using enhanced convolution neural network with attention mechanism // Computer Systems Science and Engineering. Том 41. – 2022. - №1. – С. 415-426.

17. Shahin I., Hindawi N., Nassif A.B., Alhudhaif A., Polat K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition // Expert Systems with Applications. Том 188. – 2022.

18. Steven R. Livingstone, Frank A. Russo. 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. Том 13. – 2018. - №5.

19. Andersson H.W., Lilleeng S.E., Ruud T., Ose S.O Suicidal ideation in patients with mental illness and concurrent substance use: analyses of national census data in Norway /. // BMC Psychiatry. 2022. Том 22. №1.

20. Sun X., Su K., Fan C. VFL—A deep learning-based framework for classifying walking gaits into emotions // Neurocomputing. Том 473. – 2022. – С. 1-13.

21. Kosendiak A., Król M., Ściskalska M., Kepinska M. The Changes in Stress Coping, Alcohol Use, Cigarette Smoking and Physical Activity during COVID-19 Related Lockdown in Medical Students in Poland // International Journal of Environmental Research and Public Health. 2022. Том 19. №17.

22. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). 2018. URL: https://zenodo.org/record/1188976 (дата обращения: 25.04.2022)

23. Thirumuru R., Gurugubelli K., Vuppala A.K. Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition // Digital Signal Processing: A Review Journal. Том 120. – 2022.

24. Van L.T., Nguyen Q.H., Le T.D.T. Emotion recognition with capsule neural network // Computer Systems Science and Engineering. Том 41. – 2022. - №3. – С. 1083-1098.

25. Varun Chand H., Karthikeyan J. Cnn based driver drowsiness detection system using emotion analysis // Intelligent Automation and Soft Computing. 2022. Том 31. №2. С. 717-728.

26. World Health Organization about depression. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/depression (дата обращения: 21.12.2021).

27. Yu W., Xu H. Co-attentive multi-task convolutional neural network for facial expression recognition // Pattern Recognition. Том 123. – 2022.

28. Zhu Q., Mao Q., Jia H., Noi O.E.N., Tu J. Relation network for facial expression recognition in the wild with few-shot learning Convolutional // Expert Systems with Applications. Том 189. – 2022.

References

1. Information about road safety situation in Russian Federation. URL: http://stat.gibdd.ru (request date: 21.12.2021).

2. Fei Z., Yang E., Yu L., Li X., Zhou H., Zhou W. A Novel deep neural network-based emotion analysis system for automatic detection of mild cognitive impairment in the elderly // Neurocomputing. 2022. Том 468. С. 306-316.

3. Abdulmohsin H.A., Wahab H.B.A., Hossen A.M.J.A. A novel classification method with cubic spline interpolation // Intelligent Automation and Soft Computing. Том 31. – 2022. - №1. – С. 339-355.

4. Fei Z., Yang E., Li D.D.-U., Butler S., Ijomah W., Li X., Zhou H. Deep convolution network based emotion analysis towards mental health care // Neurocomputing. 2020. Том 388. С. 212-227.

5. Almeida J., Vilaça L., Teixeira I.N., Viana P. Emotion identification in movies through facial expression recognition // Applied Sciences (Switzerland). 2021. Том 11. №15.

6. Tu G., Wen J., Liu H., Chen S., Zheng L., Jiang D. Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models[Formula presented] // Knowledge-Based Systems. 2022. Том 235.

7. GBD Result Tool search result for depression in Russian Federation. URL: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/4ba5e9f86007614b6cc22ce8b7d41f7f (дата обращения: 21.12.2021)

8. Gan C., Xiao J., Wang Z., Zhang Z., Zhu Q. Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention // Image and Vision Computing. Том 117. – 2022.

9. He L., Jiang D., Sahli H. Automatic Depression Analysis Using Dynamic Facial Appearance Descriptor and Dirichlet Process Fisher Encoding // IEEE Transactions on Multimedia. 2019. Том 21. №6. С. 1476-1486.

10. Li J., Qin D. The mutation seagull algorithm optimizes the speech emotion recognition of BP neural network // ACM International Conference Proceeding Series. 2021. С. 160-164.

11. Lyu L., Zhang Y., Chi M.-Y., Yang F., Zhang S.-G., Liu P., Lu W.-G. Spontaneous facial expression database of learners' academic emotions in online learning with hand occlusion // Computers and Electrical Engineering. Том 97. – 2022.

12. Mehrabian A. Nonverbal Communication. New York, 1972. 235 с.

13. Nair P., Subha V., Aneesh R. Non Verbal Behaviour Analysis for Distress Detection Using Texture Analysis // 2018 International CET Conference on Control, Communication, and Computing, IC4 2018. 2018. С. 239-244.

14. Pandey S.K., Shekhawat H.S., Prasanna S.R.M. Attention gated tensor neural network architectures for speech emotion recognition // Biomedical Signal Processing and Control. 2022. Том 71.

15. Parra-Gallego L.F., Orozco-Arroyave J.R. Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments // Digital Signal Processing: A Review Journal. Том 120. – 2022.

16. Prabhu K., SathishKumar S., Sivachitra M., Dineshkumar S., Sathiyabama P. Facial expression recognition using enhanced convolution neural network with attention mechanism // Computer Systems Science and Engineering. Том 41. – 2022. - №1. – С. 415-426.

17. Shahin I., Hindawi N., Nassif A.B., Alhudhaif A., Polat K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition // Expert Systems with Applications. Том 188. – 2022.

18. Steven R. Livingstone, Frank A. Russo. 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. Том 13. – 2018. - №5.

19. Andersson H.W., Lilleeng S.E., Ruud T., Ose S.O. Suicidal ideation in patients with mental illness and concurrent substance use: analyses of national census data in Norway // BMC Psychiatry. 2022. Том 22. №1.

20. Sun X., Su K., Fan C. VFL—A deep learning-based framework for classifying walking gaits into emotions // Neurocomputing. Том 473. – 2022. – С. 1-13.

21. Kosendiak A., Król M., Ściskalska M., Kepinska M. The Changes in Stress Coping, Alcohol Use, Cigarette Smoking and Physical Activity during COVID-19 Related Lockdown in Medical Students in Poland // International Journal of Environmental Research and Public Health. 2022. Том 19. №17.

22. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). 2018. URL: https://zenodo.org/record/1188976 (дата обращения: 25.04.2022)

23. Thirumuru R., Gurugubelli K., Vuppala A.K. Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition // Digital Signal Processing: A Review Journal. Том 120. – 2022.

24. Van L.T., Nguyen Q.H., Le T.D.T. Emotion recognition with capsule neural network // Computer Systems Science and Engineering. Том 41. – 2022. - №3. – С. 1083-1098.

25. Varun Chand H., Karthikeyan J. Cnn based driver drowsiness detection system using emotion analysis // Intelligent Automation and Soft Computing. 2022. Том 31. №2. С. 717-728

26. World Health Organization about depression. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/depression (дата обращения: 21.12.2021).

27. Yu W., Xu H. Co-attentive multi-task convolutional neural network for facial expression recognition // Pattern Recognition. Том 123. – 2022.

28. Relation network for facial expression recognition in the wild with few-shot learning / Zhu Q., Mao Q., Jia H., Noi O.E.N., Tu J. Convolutional // Expert Systems with Applications. Том 189. – 2022.

Демин Олег Дмитриевич, студент 2-го курса магистратуры, инженер Национального центра когнитивных разработок Университета ИТМО
Лаптев Андрей Александрович, аспирант, инженер Национального центра когнитивных разработок Университета ИТМО

Demin Oleg Dmitrievich, 2nd year Master's student, engineer at National Center for Cognitive Research of ITMO University
Laptev Andrey Aleksandrovich, PhD student, engineer at National Center for Cognitive Research of ITMO University