УДК 004.93 **DOI** 10.15622/ia.21.6.2

А.А. Двойникова, М.В. Маркитантов, Е.В. Рюмина, М.Ю. Уздяев, А.Н. Величко, Д.А. Рюмин, Е.Е. Ляксо, А.А. Карпов

АНАЛИЗ ИНФОРМАЦИОННОГО И МАТЕМАТИЧЕСКОГО ОБЕСПЕЧЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ АФФЕКТИВНЫХ СОСТОЯНИЙ ЧЕЛОВЕКА

Двойникова А.А., Маркитантов М.В., Рюмина Е.В., Уздяев М.Ю., Величко А.Н., Рюмин Д.А., Ляксо Е.Е., Карпов А.А. **Анализ информационного и математического обеспечения для распознавания аффективных состояний человека.**

Аннотация. В статье представлен аналитический обзор исследований в области аффективных вычислений. Это направление является составляющей искусственного интеллекта, и изучает методы, алгоритмы и системы для анализа аффективных состояний человека при его взаимодействии с другими людьми, компьютерными системами или роботами. В области интеллектуального анализа данных под аффектом подразумевается проявление психологических реакций на возбуждаемое событие, которое может протекать как в краткосрочном, так и в долгосрочном периоде, а также иметь различную интенсивность переживаний. Аффекты в рассматриваемой области разделены на 4 вида: аффективные эмоции, базовые эмоции, настроение и аффективные расстройства. Проявление аффективных состояний отражается в вербальных данных и невербальных характеристиках поведения: акустических и лингвистических характеристиках речи, мимике, жестах и позах человека. В обзоре приводится сравнительный анализ существующего информационного обеспечения для автоматического распознавания аффективных состояний человека на примере эмоций, сентимента, агрессии и депрессии. Немногочисленные русскоязычные аффективные базы данных пока существенно уступают по объему и качеству электронным ресурсам на других мировых языках, что обуславливает необходимость рассмотрения широкого спектра дополнительных подходов, методов и алгоритмов, применяемых в условиях ограниченного объема обучающих и тестовых данных, и ставит задачу разработки новых подходов к аугментации данных, переносу обучения моделей и адаптации иноязычных ресурсов. В статье приводится описание методов анализа одномодальной визуальной, акустической и лингвистической информации, а также многомодальных подходов к распознаванию аффективных состояний. Многомодальный подход к автоматическому анализу аффективных состояний позволяет повысить точность распознавания рассматриваемых явлений относительно одномодальных решений. В обзоре отмечена тенденция современных исследований, заключающаяся в том, что нейросетевые методы постепенно вытесняют классические детерминированные методы благодаря лучшему качеству распознавания состояний и оперативной обработке большого объема данных. В статье рассматриваются методы анализа аффективных состояний. Преимуществом использования многозадачных иерархических подходов является возможность извлекать новые типы знаний, в том числе о влиянии, корреляции и взаимодействии нескольких аффективных состояний друг на друга, что потенциально влечет к улучшению качества распознавания. Приводятся потенциальные требования к разрабатываемым системам анализа аффективных состояний и основные направления дальнейших исследований.

Ключевые слова: аффективные состояния, аффективные вычисления, эмоции, сентимент, депрессия, агрессия, корпуса данных, компьютерные системы.

1. Введение. На протяжении последнего десятилетия активно развивается направление аффективных вычислений в области искусственного интеллекта. Аффективные вычисления (англ. Affective Computing) – область искусственного интеллекта, изучающая методы, алгоритмы, системы и устройства для анализа человеческих аффектов при его взаимодействии с другим человеком или машиной (роботом) [1]. Аффективные состояния играют значительную роль в человеческой жизни. С помощью них человек осуществляет субъективную оценку явлений, событий и ситуаций; побуждается к действиям; адаптируется к ситуациям; выражает свое внутреннее состояние; осуществляет регуляцию поведения [2]. Первое определение аффекта было опубликовано в 1897 немецким психологом Крафт-Эбингом. Он считал, что аффект – это сильное душевное волнение [3]. В современной психологии термины «аффект» и «эмоции» являются связными понятиями. Считается, что аффект – сильно выраженное эмоциональное проявление, и как разновидность эмоции характеризуется быстрым возникновением, высокой интенсивностью, кратковременностью, безотчетностью [4]. В то же время некоторые психологи [5] разделяют понятия аффектов и эмоций, подразумевая под аффектами неуправляемые переживания, когда проявление эмоций – самоуправляемое состояние человека.

Определение аффекта в области аффективных вычислений несколько отличается от его дефиниций в психологии и криминалистике. В существующих исследованиях нет четкого определения аффекта, поэтому существует необходимость раскрыть термин «аффект» в области аффективных вычислений. Под аффектом подразумевается проявление психологических реакций на возбуждаемое событие, которое может протекать как в краткосрочном, так и в долгосрочном периоде, а также иметь различную интенсивность переживаний. Можно выделить несколько видов аффекта, которые обобщенно представлены на рисунке 1.

Аффекты в области интеллектуального анализа данных делятся на 4 вида: аффективные эмоции, базовые эмоции, настроение и аффективные расстройства. Аффективные эмоции (например, агрессия) характеризуются высокой интенсивностью проявлений, являются незамедлительной и неконтролируемой реакцией на события, а также имеют небольшую длительность проявлений. Агрессия – целенаправленное наступательное поведение, которое наносит вред одушевленным и неодушевленным объектам нападения [6]. Базовые эмоции (радость, грусть, удивление и др.), в отличие от аффективных, имеют более низкую интенсивность и высокую

продолжительность переживаний [7]. Настроение представляет собой эмоциональное состояние, которое может продолжаться длительный период. Автоматический анализ настроений можно наблюдать в задачах сентимент-анализа [8] и распознавания токсичности. Сентимент – субъективное выражение мнений, взглядов, отношений к ситуациям или предметам. Иногда под сентиментом подразумевают определение полярности эмоции (позитив и негатив) [7]. Под токсичностью подразумевается поведение человека, которое вносит деструктив с отрицательной полярностью в нормы общения между людьми. Как правило, токсичность проявляется в текстовой модальности, поэтому задача распознавания токсичности в комментариях социальных сетей на сегодняшний день является актуальной (https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge – 2018). Аффективные расстройства выражаются в психологических нарушениях, которые характеризуются несознательной сменой настроения, преимущественно в сторону отрицательной полярности, которое может выражаться в угрозах, нецензурной лексике, оскорблениях и пр. Как правило, аффективные расстройства, например депрессия, могут протекать от недель до нескольких лет [9]. Депрессия является нарушением психического здоровья человека, оказывающая негативное влияние на его мысли и поступки [10]. Различные аффективные состояния могут иметь корреляционную зависимость между собой.

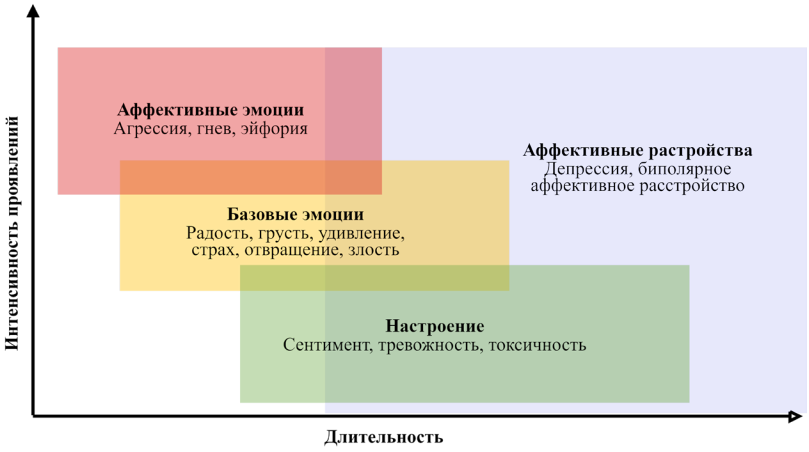

Рис. 1. Систематизация видов аффекта в области аффективных вычислений

Автоматическое определение аффективных состояний человека является актуальной и востребованной задачей в области искусственного интеллекта. Автоматические системы анализа и непрерывного мониторинга состояния человека позволяют оценивать и интерпретировать проявленные эмоции, настроения, тем самым, предоставляя возможность своевременно реагировать на изменение психоэмоциональных характеристик пользователей и адаптировать стратегию взаимодействия. В частности, использование автоматизированных экспертных систем мониторинга, распознавания и анализа состояний депрессии значительно облегчит работу врачей-специалистов и позволит получить дополнительные данные, которые могут быть недоступны из-за отсутствия постоянного контакта между пациентом и специалистом. Распознавание агрессии имеет большое значение для регуляции социальных взаимодействий не только во время живого общения, но и в виртуальном интернет-пространстве. С помощью анализа текстовой модальности можно распознавать проявления кибербуллинга и токсичности. Это особенно важно в условиях дистанционного образования, общения в социальных сетях, просмотре новостей и т.д.

Проявление аффективных состояний человека выражаются в его акустических и лингвистических характеристиках речи, а также мимики, позах, жестах, и физиологических сигналах (пульс, давление, электропроводность, цвет кожи и т.д.). При автоматическом определении аффективных состояний важно учитывать как вербальные, так и невербальные характеристики, на которые обращают внимание специалисты при личной беседе с людьми. К вербальным относятся ключевые слова, фразы и выражения, отражающие мысли говорящего, к невербальным признакам – просодические характеристики голоса (тон, тембр, громкость, интонация), мимика, направление взгляда, жесты и т.д. Для анализа физиологических сигналов используют контактные сенсоры, такие как пульсометры, электроды и пр. При наличии отвлекающих факторов (приборы на теле) у человека проявление аффективных состояний будет изменено, тем самым, анализ естественных внутренних переживаний человека становится невозможным. Поэтому репрезентативными модальностями для построения автоматической системы анализа аффективных состояний являются видео, аудио и текстовая информация. Анализ нескольких каналов информации (многомодальный подход) имеет преимущества перед одномодальным анализом [11], в том числе более высокую точность и устойчивость к

пропущенным данным, например, в случае технического сбоя или чрезмерного шума в одной из модальностей.

Целью статьи является сравнительный аналитический обзор существующих методов распознавания различных аффективных состояний человека – эмоций, сентимента, агрессии, депрессии. Для достижения цели необходимо выполнить основную задачу, которая заключается в сравнительном анализе корпусов данных, содержащих рассматриваемые аффекты. Предложенный аналитический обзор помогает выявить наиболее репрезентативное информационное обеспечение, выделить достоинства и недостатки применяемых методов для распознавания аффективных состояний, обнаружить нерешенные проблемы в рассматриваемой области, а также определить требования к разрабатываемым системам для анализа аффектов. Существует большое количество зарубежных работ на тему анализа аффективных состояний человека, однако в России количество таких исследований крайне мало. Поэтому предложенный аналитический обзор является актуальным.

2. Анализ информационного обеспечения. В мире накоплен большой объем данных (информационного обеспечения) для анализа аффективных состояний. В данной статье рассматриваются только многомодальные корпуса, ниже приводится их описание, сгруппированное по аффективным состояниям, прилагаются также сравнительные таблицы с основными характеристиками корпусов.

2.1. Эмоции. Аудиовизуальный корпус Acted Facial Expressions in the Wild (AFEW) [12] (https://cs.anu.edu.au/few/AFEW.html) содержит короткие эмоциональные записи из различных художественных фильмов (возраст актеров – 1-70 лет). Корпус аннотирован на уровне фраз. Часть записей из корпуса AFEW и собранные новые записи аннотированы на уровне кадров на пространственные оценки валентности и интенсивности и 68 координат ключевых точек лица, что позволило собрать еще один независимый корпус AFEW-VA [13] (https://ibug.doc.ic.ac.uk/resources/afew-va-database/) (возраст актеров – 8-76 лет, 52% – женщины.). Количественные оценки валентности и интенсивности измеряются в диапазоне [-10; 10]. Все данные корпуса AFEW-VA размечали 2 эксперта.

Аудиовизуальный корпус Aff-Wild2 [14] (https://ibug.doc.ic.ac.uk/resources/aff-wild2/) содержит записи (39% дикторов – женщины) YouTube. Тематика записей корпуса различная. Разметка корпуса выполнялась 4 экспертами на уровне кадра по эмоциям и единицам действия лица (Action Units) [15].

Многомодальный корпус Interactive Emotional Dyadic Motion Capture (IEMOCAP) [16] (https://sail.usc.edu/iemocap/iemocap_release.htm) собран посредством диадического взаимодействия двух актеров (5 пар актеров, 50% – женщины). Для захвата движений мышц лица, головы и рук при различных эмоциях на теле актеров наносились 59 маркеров. Каждая запись аннотирована не менее 3 экспертами на уровне высказываний, а также самооценкой.

Многомодальный корпус RECOLA [17] (https://diuf.unifr.ch/main/diva/recola/download.html) записывался посредством взаимодействия двух участников (23 группы, 59% участников – женщины) с использованием оборудования для анализа электрокардиограмм (ЭКГ) и электродермальной активности. Записи корпуса размечались 6 экспертами на уровне кадров.

Многомодальный корпус SEWA [18] (https://db.sewaproject.eu/) записывался в натурных условиях в результате двух экспериментов над участниками (возраст – 18-60 лет, 49% – женщины). В рамках первого эксперимента пара участников просматривала 4 рекламных ролика по 60 с, во время второго – эти же участники обсуждали их. Каждый участник по завершении экспериментов заполнял отчет о своем эмоциональном состоянии. Корпус аннотирован на ключевые точки лица, единицы действия лица, низкоуровневые дескрипторы речи, а также пространственные оценки валентности, интенсивности, симпатии/антипатии, согласия/несогласия и наличия случаев непредвиденных ситуаций, разметка выполнена 5 экспертами. Участники общались между собой на 6 различных языках: китайский, английский, немецкий, греческий, венгерский, сербский.

Аудиовизуальный корпус SEMAINE [19] (https://semaine-db.eu/) записывался при взаимодействии пользователей (возраст 22-60 лет, 62% – женщины) с 4 аватарами. Для записи использовались 5 видеокамер и 4 микрофона.

Аудиовизуальный корпус RAMAS [20] записан в результате взаимодействия двух актеров (5 пар, 50% – женщины, возрастной диапазон 18-20 лет). Корпус размечен 21 экспертом на уровне временных сегментов записи. В том числе актеры также предоставили самооценки их эмоций по шкале Лайкерта (англ. Likert scale) (от 0 – «нет признаков эмоции» до 3 – «сильно выражена эмоция»).

Многомодальный корпус Multimodal EmotionLines Dataset (MELD) [21] (https://affective-meld.github.io/) содержит короткие эмоциональные высказывания из телесериала «Друзья». Корпус размечен 3 экспертами отдельно на основе текстовой и

аудиовизуальной информации на уровне высказываний. Итоговая метка для данных выбиралась посредством мажоритарного голосования.

Многомодальный корпус CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [22] (https://github.com/A2Zadeh/CMU-MultimodalSDK) содержит видеоролики из YouTube. Тематика записей корпуса различна, всего рассмотрено 250 тем (обзоры, дебаты, консультации и др.). Деление на записи выполнялось с учетом начала и конца предложения. На одного диктора (43% дикторов – женщины) приходится не менее 10, и не более 50 высказываний. Каждое предложение аннотировано на 7 классов сентимента и 6 базовых категорий эмоций по шкале Лайкерта.

2.2. Сентимент. Для задачи распознавания сентимента известны 4 многомодальных корпуса данных. Корпус Multimodal Opinion Utterances Dataset (MOUD) [23] (http://multicomp.cs.cmu.edu/resources/moud-dataset/) содержит видеоролики с YouTube на различные темы. Поскольку дикторы (возраст – 20-60 лет, 81% – женщины) могли разговаривать на разные темы в одном видео, авторы корпуса выбирали 30 секундный фрагмент из каждого видео. Каждый видеофрагмент разделен на высказывания (в общей сложности 498). Разметка проводилась 2 экспертами на уровне высказываний.

Многомодальный корпус Multimodal Opinion-Level Sentiment Intensity (MOSI) [24] (http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/) представляет собой набор видеороликов из YouTube на различные темы (возраст участников – 20-30 лет, 46% – женщины). Из всех видео выбраны высказывания (в общей сложности 2199), в которых присутствовала субъективность, выражающаяся во мнении или отношении человека к чему-либо. Корпус размечен 5 экспертами на уровне высказываний по сентименту, а также по мимике, движениям лица и головы.

Авторы базы данных YOUTUBE Dataset [25] (http://multicomp.cs.cmu.edu/resources/youtube-dataset-2/) отбирали видео (возраст участников 14-60 лет, 42% – женщины) по следующим ключевым словам: мнение, обзор, лучшие духи, я ненавижу, мне нравится и др. Авторы корпуса сокращали каждое видео до 30 с, затем полученные фрагменты размечались тремя экспертами.

Китайский многомодальный корпус CH-SIMS Dataset [26] (https://github.com/thuiar/MMSA) содержит данные (34% дикторов – женщины) из различных фильмов, сериалов, телешоу. Авторы корпуса разделили каждое видео на фрагменты длиной от 1 до 10 с, в каждом

фрагменте присутствует только один диктор. Разметка корпуса производилась 5 экспертами на уровне одной записи.

Текстовая модальность является наиболее репрезентативной для определения сентимента. Известны несколько русскоязычных лингвистических корпусов [8]: RuTweetCorp, РОМИП 2012, RuSentiment, LinisCrowd, SentiRuEval, Auto_reviews и др.

Поскольку эмоции и сентимент являются родственными понятиями в области искусственного интеллекта, в таблицах 1 и 2 представлены сравнительные характеристики корпусов для распознавания эмоций и сентимента. В таблицах приводится параметр – тип речи (подготовленная или спонтанная), обозначающий вариативность лексического содержания высказываний. Во всех корпусах, где присутствует текстовая модальность данных, авторы извлекли орфографические транскрипции вручную.

Таблица 1. Характеристики многомодальных корпусов для анализа эмоций

Название корпуса	Модальности				Объем корпуса		Разметка	Количество дикторов	Условия записи	Тип речи	Язык речевых данных
	В	А	Т	Ф	Часов	Записей					
AFEW [12]	+	+	-	-	2,5	1645	7 эмоций	330	Нат	П	Английский
AFEW-VA [13]	+	+	-	-	н/д	600	валентность, интенсивность	240	Нат	П	Английский
AffWild2 [14]	+	+	-	-	43	558	7 эмоций, валентность, интенсивность	458	Нат	С	Английский
IEMOCAP [16]	+	+	+	-	11,5	10039	5 эмоций, валентность, активация, доминирование	10	Лаб	П и С	Английский
RECOLA [17]	+	+	-	+	4	46	валентность, интенсивность; доминирование; согласие, вовлеченность, исполнительность, взаимопонимание.	46	Лаб	С	Французский

Продолжение Таблицы 1

SEWA [18]	+	+	+	-	44	1990	валентность, интенсивность	398	Нат	С	6 языков
SEMAINE [19]	+	+	-	-	6,5	80	7 эмоций, валентность, активация; влияние, предвкушение; эпистемо-логические состояния, процесс взаимодействия, достоверность.	20	Лаб	С	Английский
RAMAS [20]	+	+	-	+	7	581	7 эмоций, доминирование, подчинение	10	Лаб	С	Русский

Таблица 2. Характеристики многомодальных корпусов для анализа эмоций и сентимента

Название корпуса	**Модальности**			**Объем корпуса**		**Разметка**	**Количество дикторов**	**Язык речевых данных**
	В	**А**	**Т**	**Часов**	**Записей**			
MELD [21]	+	+	+	н/д	13708	7 эмоций, 3 сентимента	6	Английский
CMU-MOSEI [22]	+	+	+	66	23453	7 эмоций, 7 сентиментов	1000	Английский
MOUD [23]	+	+	+	0,6	80	3 сентимента	80	Испанский
MOSI [24]	+	+	+	н/д	93	7 сентиментов	81	Английский
YOUTUBE [25]	+	+	+	0,3	47	3 сентимента	47	Английский
CH-SIMS [26]	+	+	+	н/д	2281	5 сентиментов	2281	Китайский

В таблице 2 все условия записи данных корпуса являются натурными, тип речи – спонтанный. Как правило, анализируется такой набор из 5-7 базовых эмоций: злость, отвращение, страх, печаль, радость, удивление и нейтральное состояние. При упоминании в таблице трех классов сентимента подразумевают позитивный, нейтральный, негативный и также могут добавляться их градации.

В таблицах 1, 2, 3 и 4 используются следующие сокращения: В – видео, А – аудио, Т – текст, Ф – физиологические сигналы, Нат – натурные, Лаб – лабораторные, П – подготовленная, С – спонтанная, н/д – нет данных.

Из таблицы 1 и 2 следует, что существует большое количество многомодальных корпусов для распознавания эмоций. Наборов данных для распознавания сентимента в различных модальностях значительно меньше, это может быть связано с тем, что большинство людей выражает свое мнение только с помощью текстовых сообщений. Также большинство корпусов имеет натурные условия записи (англ. in-the-wild) и спонтанную речь, что является значительным преимуществом для автоматической системы анализа аффективных состояний.

Помимо многомодальных корпусов для распознавания агрессии существует небольшое число одномодальных визуальных корпусов, в которых содержатся записи уличных драк с камер наблюдения [31–33], а также текстовых, представленных в работах [34, 35]. В таблице 3 представлены сравнительные характеристики многомодальных корпусов для распознавания агрессии. Во всех рассмотренных данных тип речи – спонтанный. Авторы всех рассмотренных корпусов извлекли текстовые транскрипции ручным методом. Запись корпусов TR и SD проводилась в натурных условиях, а NAA – в лабораторных.

Таблица 3. Характеристики многомодальных корпусов для анализа агрессии

Название корпуса	**Модальности**				**Объем корпуса**		**Разметка**	**Количество дикторов**	**Язык речевых данных**
	В	**А**	**Т**	**Ф**	**Часов**	**Записей**			
TR [27]	+	+	+	-	0,6	н/д	3 уровня агрессии	н/д	Нидерландский
SD [28, 29]	+	+	+	-	0,5	8	3 уровня агрессии 5 уровней стресса	9	Английский, нидерландский
NAA [30]	+	+	+	+	н/д	2240	5 уровней агрессии, страха, интенсивности 9 уровней валентности	16	Нидерландский

Анализ использованных методик записи корпусов показывает слабую проработку критериев определения агрессивного поведения при разметке: для оценки агрессивного поведения часто агрессия путается с гневом, демонстративным поведением и т.д. Все рассмотренные корпуса, содержащие многомодальные проявления агрессии, находятся в закрытом доступе. Кроме этого, данные корпуса имеют малые размеры и малую репрезентативность с точки зрения количества представленных людей. Ввиду этого остро стоит необходимость сбора и разметки нового корпуса многомодальных данных, содержащих агрессивное поведение.

В дальнейшем авторами данной статьи планируется собрать корпус данных, содержащий агрессивное поведение пользователей информационного пространства в различных модальностях: двигательная активность и мимические изменения, представленные на видео, невербальное речевое поведение, представленное в аудиосигнале, вербальное речевое поведение в текстовых транскрипциях. Другими отличительными особенностями данного корпуса являются большая по сравнению с известными корпусами репрезентативность испытуемых, отбор поведения русскоязычных пользователей, а также относительная естественность поведения, которая обеспечивается тем, что в данный корпус планируется отбор поведения пользователей в ходе прямых трансляций в сети Интернет, что полностью исключает вмешательство внешнего наблюдателя в поведение испытуемых.

2.3. Депрессия. Корпус Distress Analysis Interview Corpus (DAIC) [36] содержит записи клинических интервью, а целью его разработки было определение наличия риска посттравматического стрессового расстройства (ПТСР) и большого депрессивного расстройства (БДР). В корпусе содержатся несколько типов интервью: очные, телеконференции, интервью в режиме «Волшебник страны Оз» (с использованием анимированного виртуального интервьюера Элли, которая управлялась интервьюером в другой комнате), автоматические интервью (с использованием Элли в автоматическом режиме без управления интервьюером в другой комнате).

В работе [37] представлен многомодальный корпус Audio-Visual Depressive language corpus (AViD-Corpus), который включает в себя записи взаимодействия людей с компьютером. Информантам были даны следующие задания: чтение новелл и преданий, пение и др. Корпус был аннотирован на уровне записей согласно шкале Бека-2 (это второй пересмотр опросника Бека, который был принят в 1996 году). Опросник состоит из 21 вопроса, а каждый вопрос оценивается

по шкале от 0 до 3 в зависимости от тяжести симптомов. Общий балл составляет от 0 до 63, где чем выше значение, тем серьезнее симптомы депрессии.

Корпус Pitt [38] был собран на основе клинического интервью с использованием шкалы Гамильтона (англ. The Hamilton Depression Rating Scale, HRSD) и опросника симптомов депрессии (англ. The 16-item Quick Inventory of Depressive Symptomatology and Self-Report, QIDS-SR). Шкала депрессии Гамильтона является одним из наиболее часто используемых инструментов для выявления симптомов депрессии, а также считается стандартом оценки эффективности медикаментозного лечения депрессивных расстройств. Заполняется специалистом при проведении клинического интервью, общий балл составляет от 0 до 21, где чем выше значение, тем серьезнее симптомы депрессии. QIDS-SR состоит из 16 вопросов для самооценки депрессивных симптомов, имеет также вариации для специалистов (QIDS-C), является упрощенной версией опросника Inventory of Depressive Symptoms (IDS), состоящего из 30 вопросов. Суммарный балл составляет от 0 до 27.

Многомодальный корпус BlackDog [39] содержит записи интервью, которые состояли из вопросов с открытым ответом. Бинарные метки классификации («здоровые субъекты» и «тяжелая депрессия») были проставлены вручную.

Корпус Sonde Health Free Speech (SH2-FS) содержит записи информантов в естественных условиях (в машине, дома, на работе), представлен в работе [40]. Аннотация сделана с использованием данных самодиагностического теста PHQ-9. Данный тест также является одним из часто используемых для самооценки симптомов депрессии, состоит из 9 вопросов. Суммарный балл составляет от 0 до 27.

Для распознавания депрессии существуют также одномодальные корпуса [9]: аудиокорпус Mundt, русскоязычный текстовый корпуса RusNeuroPsych, корпус эссе, а также корпус информации из профилей социальной сети “ВКонтакте” (изображения, собранные из альбомов, аватаров, постов в профилях). В таблице 4 представлена сравнительная характеристика многомодальных корпусов для распознавания депрессии. Запись всех рассмотренных корпусов проводилась в натурных условиях. В корпусе DAIC транскрипции извлечены автоматическим способом, а в корпусе Pitt – ручным. Во всех корпусах язык речевых данных – английский.

Во многих случаях исследователи сталкиваются с проблемой нехватки данных из-за того, что речевых корпусов, содержащих

проявления депрессии, относительно мало, что естественно ввиду множества факторов. Например, область паралингвистики является относительно новой, хотя к ней проявляется повышенный интерес среди ученых. Кроме того, процесс сбора таких специфичных данных является трудоемким и времязатратным, а также не всегда возможно провести запись в естественных условиях, что непосредственно сказывается на количестве существующих речевых корпусов, их объемах и длительности аудиозаписей.

Существующие корпуса для задач аффективных вычислений являются ограниченными по размеру и количеству доступных данных. Русскоязычные корпуса значительно уступают базам данных на других мировых языках, таких как английский, немецкий или китайский, что обуславливает необходимость в рассмотрении широкого спектра дополнительных методов и алгоритмов решения задачи автоматического распознавания в условиях ограниченного объема доступных данных и в том числе в разработке новых подходов к аугментации данных, переносу обучения и адаптации иноязычных ресурсов**.**

Таблица 4. Характеристики многомодальных корпусов для анализа депрессии

Название корпуса	**Модальности**			**Объем корпуса, часов**	**Разметка**	**Количество участников**	**Тип речи**
	В	**А**	**Т**				
DAIC [36]	+	+	+	73,2	5 опросников	н/д	С
Pitt [38]	+	+	+	5,9	5 уровней депрессии	19	П
AViD-Corpus [37]	+	+	-	240	4 уровня депрессии	292	П и С
SH2-FS [40]	+	+	-	16	5 уровней депрессии	887	С
BlackDog [39]	+	+	-	8,5	5 уровней депрессии	30	С

3. Анализ методов автоматического распознавания аффективных состояний. Общая схема базовой системы автоматического распознавания аффективных состояний [7] представлена на рисунке 2.

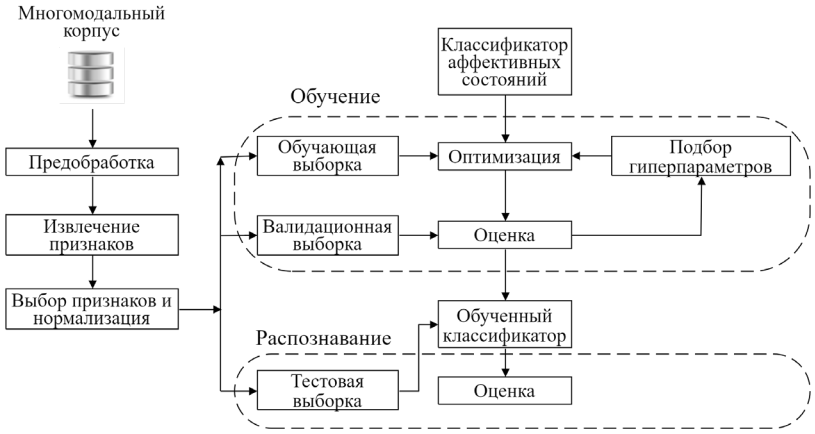

Рис. 2. Общая схема базовой системы автоматического распознавания аффективных состояний

Важными этапами в данной схеме являются извлечение информативных признаков, их выбор и нормализация, а также выбранный метод классификации. Различные методы, которые используются на данных этапах, будут рассмотрены ниже. Наиболее информативными модальностями для анализа аффективных состояний являются видео, аудио и текстовые данные. На рисунке 3 показаны основные современные методы извлечения признаков и классификации данных по каждой из модальностей; синим, желтым и фиолетовым цветами отмечена информация, относящаяся к видео, аудио и текстовой модальностям, соответственно, зеленым выделена информация, имеющая отношение ко всем модальностям одновременно.

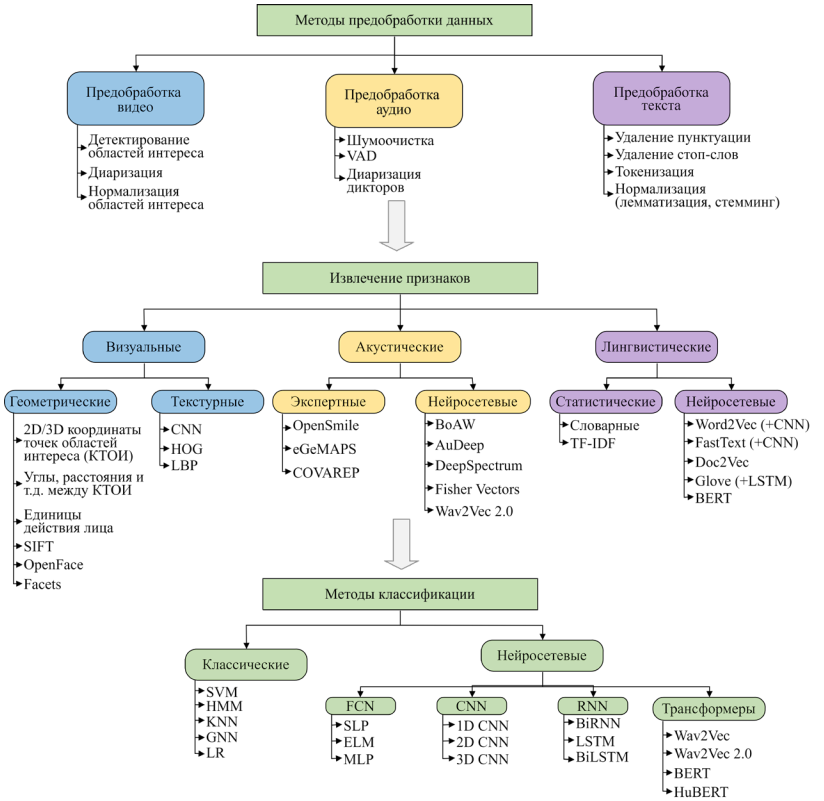

Рис. 3. Классификация методов извлечения признаков и классификации для видео, аудио и текстовой модальностей

Для каждой модальности ниже представлены описание основных методов извлечения признаков и классификации.

3.1. Видеомодальность. Классификация методов представления визуальных признаков. В данном разделе рассмотрены наиболее популярные подходы к извлечению визуальных признаков. Условно все признаки можно разделить на две группы: геометрические и текстурные признаки.

К геометрическим признакам относятся, например, 2D/3D координаты точек областей интереса, которые можно извлечь с помощью программных библиотек с открытым исходным кодом: OpenPose (https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/index.html) – детектирование 3D 135 координат ключевых точек тела, лица, рук и ног; MediaPipe

(https://google.github.io/mediapipe/) – детектирование 3D 33 координат ключевых точек тела, 21 – точек рук, 468 – точек лица; Dlib (http://dlib.net/) – детектирование 68 ключевых точек лица и др. На основе 2D/3D координаты точек областей интереса извлекаются признаки: расстояния и углы между координатами [41]; площади областей интереса [42]; единицы действия лица (action units [15]); пространственные изменения координат точек областей интереса, вычисляемые с помощью подхода Scale-Invariant Feature Transform (SIFT) [18], и др. Известны библиотеки с открытым исходным кодом, например, OpenFace (https://github.com/TadasBaltrusaitis/OpenFace) и Facets (https://imotions.com/blog/facial-expression-analysis/), извлекающие на основе геометрии лица готовые наборы признаков, включающие 2D/3D координаты точек лица, единицы действия лица и др.

Текстурные признаки извлекаются, например, с помощью сверточных нейронных сетей (англ. Convolutional Neural Network, CNN). К наиболее эффективным CNN для задачи распознавания аффективных состояний относятся архитектуры нейросетей ResNet [43] и VGG [44]. Также для извлечения текстурных признаков можно выделить подход Local Binary Patterns (LBP) [45]. При таком подходе значения каждого центрального пикселя и 8 значений пикселей в его окрестности преобразуются из значений в диапазоне от 0 до 255 в бинарные, что позволяет выделить на изображении лица углы и края, характерные для определенной мимики. Еще одним подходом для извлечения текстурных признаков является Histogram of Oriented Gradients (HOG) [46]. При таком подходе признаками являются значения градиентов (производных по x и y), которые увеличиваются по краям и углам (в областях резкого изменения интенсивности пикселей).

Для извлечения геометрических признаков требуется более надежные подходы обнаружения, так как они более чувствительны к шуму (повороту головы, окклюзия лица и т.д.). В то время как текстурные признаки трудно обобщить под всех людей, поскольку они зависимы от освещения и цвета кожи.

3.2. Аудиомодальность. Классификация методов представления акустических признаков. Акустические признаки можно разделить на два типа: экспертные и нейросетевые. Их классификация показана на рисунке 3. Ниже приводится описание акустических признаков, используемых для анализа аффективных состояний.

Экспертные/ручные (англ. hand-crafted) наборы признаков основаны на знаниях об акустических свойствах речевых сигналов. Данные признаки обычно определяются на двух уровнях: сегментов аудиосигнала (или низкоуровневые дескрипторы, англ. Low Level Descriptor, LLD) и речевого высказывания. LLD извлекаются из коротких аудиосегментов и дают мгновенную информацию об аудиосигнале. Признаки на уровне высказывания получаются путем применения статистических функционалов к полученным дескрипторам.

Низкоуровневые дескрипторы можно условно разделить на следующие категории: энергетические, просодические, вокализованные и спектральные. Каждая группа показателей предназначена для описания отдельных аспектов голоса и находит свое применение при распознавании аффективных состояний. Просодические признаки отражают особенности речи, связанные с мелодическими, временными и тембровыми характеристиками голоса, а также ритмом высказывания. Вокализованные признаки отражают отклонения частоты сигнала (джиттер) и амплитуды (шиммер), отношение гармоник частоты основного тона (ЧОТ, англ. F0) к шуму. Данные характеристики качественно влияют на восприятие голоса. Спектральные признаки характеризуют речевой сигнал в его физическом и математическом смысле на основе наличия периодической (тональной) и непериодической (шумовой) спектральных составляющих. Они позволяют отразить особенности спектра голосовых импульсов, функции голосового тракта и динамику артикуляционных органов. Энергетические признаки передают информацию об уровне энергии аудиосигнала, и эффективны для распознавания интенсивности выраженной эмоции. Они могут быть выражены через сумму спектра, частоту нуль-пересечений и т.д.

Нейросетевые признаки автоматически извлекаются алгоритмами машинного обучения и часто не могут интерпретироваться людьми. Созданные признаки могут быть определены на двух разных уровнях: сегментов аудиосигнала и высказывания.

Наборы признаков openSMILE [47] фактически являются стандартом, используемым в качестве основы в различных задачах компьютерной паралингвистики [48]. В зависимости от конфигурации, набор признаков может включать в себя от 30 до 65 низкоуровневых дескрипторов: сумма акустических спектров (громкость); мел-частотные кепстральные коэффициенты (англ. Mel-Frequency Cepstral Coefficients, MFCC) 0-14; логарифмическая мощность Mel-частотных

полос 0-7 (распределенная в диапазоне от 0 до 8 кГц); отношение гармоника/шум, джиттер, шиммер; огибающая сглаженного контура F0 с последующим сглаживанием Витерби; частота нуль-пересечений (англ. Zero-Crossing Rate) и т.д. Применение соответствующих производных к каждому дескриптору и функционалов (абсолютное положение максимального и минимального значения, стандартное отклонение, асимметрия, эксцесс, процентили 1, 25, 50, 75 и 99 %, сглаженный контур F0, локальный и дифференциальный межкадровый джиттер и т.д.) дает вектор от 1428 до 6373 признаков.

Расширенный женевский минималистический набор акустических параметров (англ. Extended Geneva Minimalistic Acoustic Parameter Set, eGeMAPS) – популярный набор признаков, разработанных вручную экспертами специально для распознавания эмоций по речи [49]. Он состоит из двух функциональных дескрипторов, среднего арифметического и коэффициента вариации набора из 42 дескрипторов.

Альтернативой вышеуказанным признакам, является использование алгоритмов машинного кодирования признаков, таких как BoAW (англ. Bag-of-Audio-Words) [50] и векторы Фишера (англ. Fisher Vectors, FV) [51]. В подходе BoAW выполняется кластеризация по 12 MFCC и логарифмической энергии сигнала. Список конечных центроидов кластера приводит к кодовой книге, которая используется для квантования исходных акустических LLD и присвоения их ближайшему (с точки зрения евклидова расстояния) звуковому слову в кодовой книге. Частоты звуковых слов используются для создания гистограммы, которая служит вектором признаков для классификации. FV обеспечивает надсегментное кодирование низкоуровневых дескрипторов, таких как MFCC и RASTA-образных коэффициентов перцептивного линейного предсказания (англ. Perceptual Linear Prediction, PLP), по их отклонению от распределения, которое можно смоделировать с помощью смеси гауссовых распределений (англ. Gaussian Mixture Model, GMM). Оба подхода успешно применяются для распознавания эмоций [50, 51].

Автоматическое извлечение признаков стало возможным с появлением нейронных сетей, способных анализировать представления признаков на ранних уровнях. Примерами являются нейросетевые наборы признаков Deep Spectrum [52], извлеченные с помощью глубокой нейронной сети, AuDeep, извлеченные с помощью архитектуры глубокой рекуррентной нейронной сети [53], и TRILL [54]. Также набирают популярность признаки, полученные с помощью Wav2Vec 2.0 [55].

Существует репозиторий алгоритмов обработки речи COVAREP (англ. Cooperative Voice Analysis Repository) [56], где исследователи могут хранить реализации опубликованных алгоритмов. Так создатели корпуса CMU-MOSEI [22] извлекли 12 MFCC, F0, функции сегментации (не)вокализованных звуков, параметры наклона пиков и коэффициенты дисперсии максимумов. Все выделенные признаки связаны с эмоциями и тоном речи.

Эффективность каждого типа признаков зависит от поставленной задачи: в то время как BoAW, Deep Spectrum и AuDeep потенциально способны превзойти функции openSMILE из-за большей гибкости и адаптивности, многие другие факторы, такие как размер выборки и шум, отрицательно влияют на производительность и эффективность обобщения. Показано, что признаки openSMILE демонстрируют более высокую производительность по сравнению с другими типами признаков на небольших наборах данных [48].

Вместо извлечения нескольких предопределенных признаков часто используется подход «грубой силы», который извлекает огромное число всевозможных признаков. В результате вектор признаков достигает размерности сотен и тысяч компонентов. Даже если предположить, что каждый признак, входящий в такой набор, является информативным, не все элементы статистически независимы друг от друга. В таком случае, а также, когда данные неинформативны и, следовательно, избыточны, возникает необходимость уменьшения размерности признаков. После данной процедуры происходит неизбежная потеря исходной информации. Различные методы в разной степени сохраняют исходную информацию; наименьших потерь можно добиться с помощью метода главных компонент (англ. Principle Component Analysis, PCA) [57].

3.3. Текстовая модальность. Классификация методов представления лингвистических признаков. В данном разделе рассматриваются наиболее часто используемые методы лингвистических векторных представлений, полученных из транскрипций естественной речи в задачах распознавания эмоций, сентимента, агрессии и депрессии. На рисунке 3 показана классификация лингвистических признаков, они подразделяются на статистические и нейросетевые.

Словарные признаки [8] представляют собой набор статистических параметров для каждого слова, полученные из тональных и морфологических словарей. TF-IDF (англ. Term Frequency - Inverse Document Frequency) – отношение частоты слова в текущем документе к частоте этого же слова во всех документах.

Наиболее часто используются нейросетевые методы представления информации, такие как Word2Vec, FastText, Glove, Doc2Vec, BERT, подробное описание данных методов представлено в работе [8]. Некоторые признаки для векторизации текста являются комбинацией из различных методов. Так, например, часто используется CNN, на вход которой подаются Word2Vec или FastText [58], а также LSTM в комбинации с Glove [59].

3.4. Классификация методов распознавания аффективных состояний. Ранние системы автоматического распознавания аффективных состояний были основаны на методах k-ближайших соседей (англ. K-Nearest Neighbor, KNN) [60], GMM [61], скрытых марковских моделях (англ. Hidden Markov Model, HMM) [62], опорных векторов (англ. Support Vector Machine, SVM) [48], логистической регрессии (англ. Logistic Regression, LR) (рисунок 3).

С появлением больших объемов данных обучение подобных систем стало времязатратным процессом. Поэтому на смену им пришли полносвязные нейронные сети (англ. Fully Connected Network, FCN), в частности, однослойный перцептрон (англ. Single-Layer Perceptron, SLP) [63], машина экстремального обучения (англ. Extreme Learning Machine, ELM) [64] и многослойный перцептрон (англ. Multi-Layer Perceptron, MLP) [65]. Затем возникли глубокие нейронные сети (англ. Deep Neural Networks, DNN), которые с каждым годом эволюционируют и становятся все более производительными [66], в частности, активно стали использоваться нейросетевые архитектуры, основанные на сверточных (англ. Convolutional Neural Networks, CNN) и рекуррентных (англ. Recurrent Neural Networks, RNN) нейронных сетях [67]. Одна из разновидностей RNN, называемая моделью с долгой краткосрочной памятью (англ. Long Short-Term Memory, LSTM), особенно популярна благодаря своей способности моделировать большие временные последовательности [68]. В результате LSTM оказался более эффективной сетью, чем простые модели RNN для распознавания эмоций [69]. Другие варианты рекуррентных сетей, такие как двунаправленные RNN и LSTM были придуманы, чтобы иметь возможность моделировать не только предыдущий, но и будущий контекст [70]. RNN также часто сочетаются с другими типами нейронных сетей, например с CNN [71]. Механизм внимания (англ. Attention) – концепция, которая может использоваться в рекуррентных нейронных сетях для улучшения механизма памяти [72]. Он позволяет фокусироваться на наиболее важных признаках и отбрасывать менее важные. Самовнимание (англ. Self-Attention) – разновидность механизма внимания, задачей которой

является выявление закономерности только между входными данными. Данная методика показала себя настолько эффективной в задаче машинного перевода, что позволила отказаться от использования RNN и заменить их на обычные нейронные сети в комбинации с механизмом самовнимания в архитектуре трансформер (англ. Transformer) [73]. Это позволило ускорить работу алгоритма, поскольку теперь каждый фрагмент может обрабатываться параллельно, в отличие от последовательной обработки в RNN. Данный механизм успешно применяется и в приложениях распознавания аффективных состояний [74]. Известной моделью-трансформером для распознавания аффективных состояний является HuBert [75]. Однако данные подходы требуют большого количества обучающих данных, поэтому в сочетании с трансформерами часто используется метод переноса обучения (англ. Transfer Learning) [76]. Идея состоит в том, чтобы научить нейронную сеть выполнять одну задачу, где данных много, а затем настроить последние слои в предобученной сети для последующей целевой задачи. Это работает, потому что ранние слои нейронных сетей обычно анализируют представления признаков низкого уровня, которые являются общими для понимания речевых сигналов в целом. Известно, что методы анализа аудиомодальности лучше работают с активацией и доминацией [77] и имеют потенциал для обобщения на разные языки, однако обычно страдают от низкой точности распознавания валентности [78]. Предобученные Wav2Vec и HuBERT позволяют справиться с данной проблемой [55]. Существуют также интегральные (англ. End-to-end, E2E) методы, основанные на нейронных сетях, которые получая на вход сырой сигнал, без предварительной обработки позволяют предсказывать аффективные состояния [79], однако они требуют большого количества данных по сравнению с классическими подходами, которые включают в себя различные шаги по предобработке аудиосигнала и извлечению признаков.

3.5. Методы многомодального объединения информации. Как было сказано выше, применение систем распознавания аффективных состояний человека с помощью только одной модальности (аудио, видео или текст) имеет ряд ограничений. Эти ограничения связаны как с техническими аспектами (неисправность камер, микрофонов и других сенсоров, высокий уровень шума и т.д.), так и с неоднозначностью интерпретации аффективных состояний исключительно по одному типу сигнала. Многочисленные источники литературы показывают, что валентность эмоции поддается моделированию с помощью акустических параметров значительно

хуже, чем активация [80]. То есть, довольно легко определить по голосу, спокоен или возбужден человек, но сложно однозначно сказать, в каком ключе: положительном или отрицательном. Такие эмоции, как счастье и злость имеют схожие акустические характеристики: повышение значения ЧОТ, высокая вариативность ЧОТ, повышение энергии в голосе и др. [7], что зачастую делает распознавание этих эмоций по голосу затруднительным. Для того, чтобы отличить эти эмоции, можно взглянуть на лицо человека и его мимику. Таким образом, одномодальный подход, использующий только один тип информации, обладает значительными ограничениями. Поведение человека складывается из речи, мимики и жестов. Восприятие этой информации у человека является многомодальным, т.е. осуществляется одновременно по нескольким каналам – визуальному, аудиальному, тактильному и т.д. Комбинирование двух и более модальностей позволяет значительно увеличить точность распознавания аффективных состояний [11], так как низкая точность работы системы классификации по одной модальности может компенсироваться высокой точностью по другой. Многомодальное распознавание также позволяет справиться с проблемой пропуска информации (когда в силу неисправности оборудования или плохого качества данных не удается использовать информацию от какой-либо модальности). Кроме того, многомодальный анализ позволяет зачастую распознавать и такие аффективные состояния, как сарказм и ирония, которые характеризуются явным несовпадением смысла высказывания (анализ текста) с интонацией (анализ аудио) и мимикой (анализ видео). Следовательно, аудиовизуальный анализ характеристик диктора может значительно улучшить эффективность существующих систем распознавания аффективных состояний диктора.

Существуют несколько основных подходов к объединению аудио, видео и текстовой информации на различных уровнях: а) признаков, б) представлений признаков, в) моделей, г) гипотез предсказаний и д) гибридные модели [81]. На рисунке 4 схематично показаны подходы к объединению двух модальностей (на примере аудио и видео), однако, все представленные подходы являются легко масштабируемыми при добавлении новых модальностей.

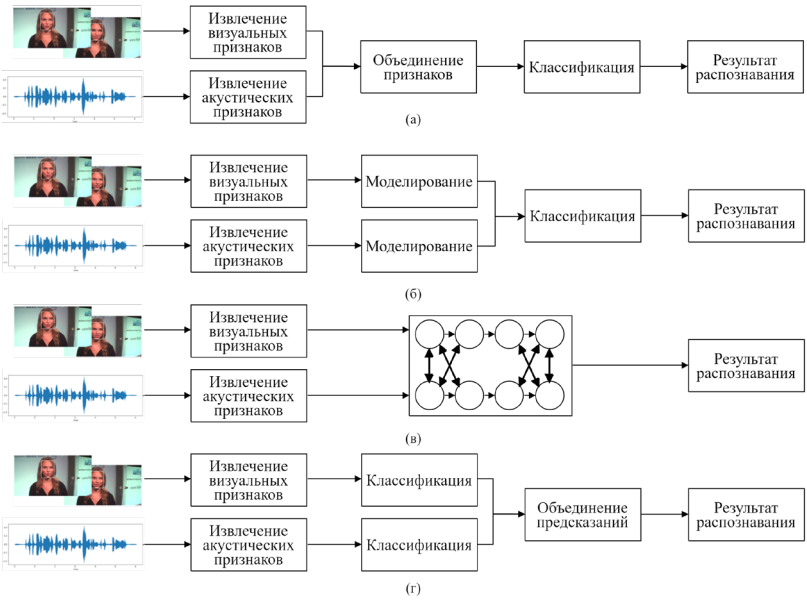

Рис. 4. Основные подходы к многомодальному объединению информации

С точки зрения моделирования, способы объединения на уровне признаков (рисунок 4а) и представлений признаков (рисунок 4б) являются наиболее предпочтительными, так как позволяют одновременно учитывать не только различные аспекты одного и того же явления, но также и их взаимосвязи между собой. Это самый естественный способ обработки информации, по принципу которого работает человеческий мозг. Однако реализация такого объединения значительно затруднена, в частности, из-за различной природы акустических, визуальных и текстовых признаков, различного распределения этих признаков, проблем, связанных с различным количеством информации на единицу времени (частота вычисления векторов признаков) и синхронизацией сигналов. Использование общего набора признаков ведет к увеличению размерности признакового пространства, что при ограниченном объеме данных может сказаться на качестве распознавания в худшую сторону [82]. Объединение информации на уровне моделей на примере использования методов HMM показано на рисунке 4в. При таком подходе создаются две различные модели, которые обмениваются информацией и выдают общее предсказание. Применение данного подхода ограничено несколькими классификационными методами,

например, некоторыми видами HMM и нейронными сетями [83], однако другие методы классификации, например SVM, не подходят в виду того, что у них отсутствует механизм, принимающий во внимание состояния других классификационных моделей. Исследования показывают, что механизмы внимания широко используются и при объединении нескольких модальностей [84]. Наиболее простым и эффективным способом объединения информации от разных модальностей является объединение на уровне предсказаний/гипотез распознавания (рисунок 4г). Объединение предсказаний возможно как методом голосования (жесткая классификации), так и с помощью суммирования и/или взвешивания вероятностей (мягкая классификация) [85]. Такой подход имеет несколько важных преимуществ, например, не существует ограничений в выборе методов классификации, и для каждой модальности возможно подобрать оптимальный метод моделирования. Однако недостатком данного подхода является то, что он не учитывает взаимосвязи между различными модальностями, предполагая, что они независимы. Гибридные методы сочетают в себе сразу несколько описанных подходов к объединению информации [86].

В современных работах в области аффективных вычислений используются как классические экспертные методы, так и нейросетевые для извлечения признаков и классификации. Тенденция исследований показывает, что нейросетевые методы постепенно вытесняют экспертные, за счет достижения большей точности распознавания аффективных состояний и быстрой обработки больших объемов данных. Многомодальный подход для автоматического анализа аффективных состояний позволяют повысить точность распознавания аффектов.

4. Методы для распознавания аффективных состояний. В данном разделе рассматриваются актуальные исследования, в которых применяются методы распознавания эмоций, сентимента и агрессии. Методы распознавания депрессии детально рассмотрены в работе [9].

4.1. Эмоции и сентимент. Для сравнительного анализа рассмотрены экспериментальные исследования, проводимые только на англоязычном многомодальном корпусе CMU-MOSEI [22]. К преимуществам данного корпуса можно отнести: разметку одновременно и по эмоциям, и по сентименту; натурные условия записи и спонтанную речь информантов; максимально доступное число участников (1000); преимущественно один участник в кадре; деление записей на полные высказывания. В различных работах

классификация сентимента проводилась по нескольким классам: 2 (негативный, позитивный), 3 (негативный, нейтральный, позитивный), 5 (сильно негативный, негативный, нейтральный, позитивный, сильно позитивный) и 7 классов (от -3 до 3), в соответствии с разметкой авторов CMU-MOSEI. В таблице 5 представлен сравнительный анализ автоматических систем распознавания эмоций и сентимента на корпусе CMU-MOSEI. В таблице действуют следующие обозначения: B – видео, A – аудио, T – текст, WAcc – взвешенная точность, WF – взвешенная F-мера, m – среднее значение WAcc и WF по 6 классам (таблица 5).

Во многих работах в качестве акустических, визуальных и лингвистических признаков использовались методы COVAREP, Facets и Glove, соответственно. Для данных корпуса CMU-MOSEI все признаки находятся в открытом доступе (https://github.com/A2Zadeh/CMU-MultimodalSDK).

Методы классификации условно можно разделить на группы в зависимости от используемых в них нейросетевых моделей: графовые нейронные сети – Graph Memory Fusion Network (Graph-MFN) [22] и Adversarial Representation Graph Fusion (ARGF) [87]; рекуррентные нейронные сети – Multi-Modal Multi-Utterance - Bi-Modal Attention (MMMU-BA) [88], Multi-task Multi-modal Emotion and Sentiment (MTMM-ES) [89], Interaction Canonical Correlation Network (ICCN) [90], Hierarchical Feature Fusion Network (HFFN) [91] и Inter-modal Interactive Module for Multi-modal Sentiment and Emotion Recognition (IIM-MMSE) [92]; трансформеры – Transformer-Based Joint-Encoding (TBJE) [93] и Multimodal Transformer (MulT) [94]. Наибольшая точность по показателю mWAcc для бинарного распознания 6 категорий эмоций достигается с помощью методов классификации на основе моделей трансформеров TBJE и MulT. Причиной этого является использование не только трансформеров, но также CNN для представления визуальных признаков, мел-спектрограммы и банк лог-фильтров энергии (англ. Log-Filter bank energy) – для акустических признаков. Кроме того, с помощью TBJE удалось достичь наибольшей точности распознавания 2 классов сентимента по показателю WAcc, а с помощью методов классификации на основе рекуррентных нейросетей IIM-MMSE [92] и ICCN [90] для 5 и 7 классов сентимента, соответственно.

Таблица 5. Сравнительный анализ автоматических систем распознавания эмоций и сентимента

Работа	**Признаки (B; A; T)**	**Метод классификации**	**Качество распознавания (%)**				
			Эмоции		**Сентимент**		
			mWAcc	**mWF**	**Число классов**	**WAcc**	**WF**
Khare A. и др. [94]	CNN; банк лог-фильтров энергии; Glove	MulT	67,4	78,6	-	-	-
Zadeh A. и др. [22]	Facets, OpenFace, CNN; COVAREP; Glove	Graph-MFN	62,3	76,3	2	76,9	77,0
					5	45,1	-
					7	45,0	-
Chauhan D.S. и др. [92]	-	IIM-MMSE	63,0	**79,0**	2	80,4	78,2
					5	**49,2**	-
					7	50,1	-
Delbrouck J.B. и др. [93]	CNN; Мел-спектрограммы; Glove	TBJE	**80,7**	76,7	**2**	**81,5**	
					7	44,4	-
Sun Z. и др. [90]	Facets; COVAREP; BERT	ICCN	-	-	**7**	**51,6**	-
Akhtar M.S. и др. [89]	Facets; COVAREP; Glove	MTMM-ES	62,8	78,6	2	80,5	78,8
Ghosal D. и др. [88]		MMMU-BA	-	-	2	79,8	-
Mai S. и др. [91]		HFFN	-	-	3	60,4	59,1
Mai S. и др. [87]		ARGF	-	-	**3**	**60,9**	**59,5**

Таким образом, большей точности распознавания эмоций и сентимента можно достичь с использованием машинной классификации на основе моделей рекуррентных нейросетей и трансформеров.

4.2. Агрессия. Ряд работ посвящен распознаванию вербальной агрессии в тексте. В работах [95, 96] рассматриваются методы распознавания агрессии на основе правил и словарей. В работе [97] сведены основные результаты классических методов машинного обучения для распознавания агрессии в тексте: логистической регрессии, случайного леса, ансамблевого классификатора, SVM, а также приводятся современные подходы глубокого обучения для распознавания агрессии в тексте. При этом в ряде работ рассматриваются также различные виды агрессии, которые проявляются в тексте. Выделяют прямую/косвенную [98] агрессию, а также различные разновидности прямой вербальной агрессии в тексте: проявления ненависти, оскорбления и сквернословия [99].

Работы, посвященные распознаванию агрессии по невербальному речевому поведению в аудиосигнале, можно разделить на те, которые рассматривают классические методы машинного обучения и методы, использующие глубокие нейронные сети. Классические методы машинного обучения отражены в работах [100], раскрывающих классификацию спектральных признаков посредством SVM. В работе [101] представлена классификация главных компонент спектра давления воздуха посредством HMM. Применение глубоких CNN, обрабатывающих мел-спектрограммы рассмотрено в работе [102].

Системы распознавания агрессии в поведенческой активности на видео разделяются на те, которые рассматривают сконструированные вручную дескрипторы и классические алгоритмы машинного обучения [31] и методы обработки видео глубокими нейронными сетями [103]. Подавляющее количество работ из данной категории направлены на распознавание агрессии людей, записанной на камеры внешнего видеонаблюдения.

Среди многомодальных методов распознавания агрессии выделяют методы, использующие в своей основе подходы онтологического моделирования вместе с классическими алгоритмами машинного обучения [27], а также подходы глубокого обучения [104]. На рисунке 5 изображена классификация методов распознавания агрессии пользователей.

Таким образом, на основе анализа приведенных работ выделяются основные группы методов и моделей распознавания агрессии. Можно выделить следующие критерии для классификации методов и моделей: 1) по модальности; 2) по виду распознаваемой агрессии; 3) по методу распознавания.

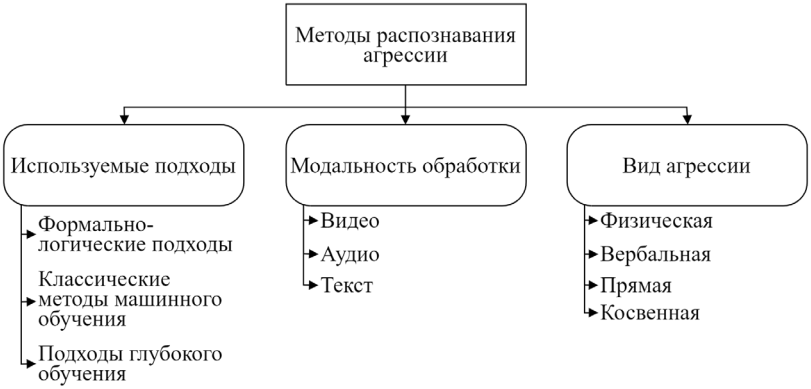

Рис. 5. Классификация методов распознавания агрессии

4.3. Многозадачные системы. Многозадачные (англ. multi-task) системы предполагают распознавание нескольких аффективных состояний. Зачастую такие системы являются и многомодальными, например, при распознавании эмоций и анализа сентимента [89]. Последний, в свою очередь, определяется, в большинстве случаев, по текстовой модальности [8], в отличие от эмоций, которые можно определить по тексту, аудио и видео. Поэтому многозадачные системы схожи с многомодальным объединением информации на уровне моделей (рисунок 4в) [89]. Существуют два основных подхода к разработке многозадачных систем классификации и регрессии: 1) использование общих параметров, 2) использование общих, и собственных параметров для каждой задачи. Архитектуры данных моделей представлены на рисунке 6. Количество выходных данных (предсказаний) увеличивается соответственно количеству классов рассматриваемых аффективных состояний, и можно одновременно решать задачи классификации и регрессии [105].

Общая часть системы обычно состоит из различных видов рекуррентных слоев, которые способны улавливать контекст. Этими слоями могут быть LSTM [106], GRU (англ. Gated Recurrent Unit) [107], BiLSTM [108] и др. Одним из ключевых этапов при разработке многозадачной системы обучения является определение целевой функции обучения, т.е. функции потерь. Многозадачная экспериментальная установка предполагает распознавание нескольких аффективных состояний, поэтому функция потери должна одинаково учитывать ошибки распознавания всех рассматриваемых задач, что обуславливает необходимость разрабатывать сложную составную

функцию потерь, учитывающую несколько факторов. Это может быть взвешенная сумма нескольких целевых функций, каждая из которых отвечает за свою задачу [107-109]. Механизм внимания также применяется в многозадачных системах [108].

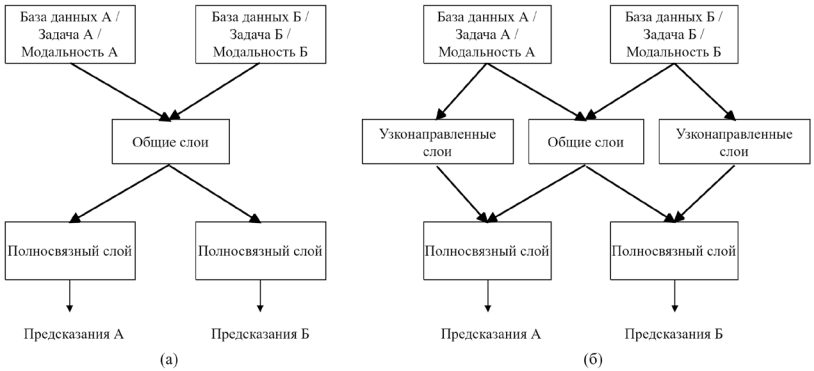

Рис. 6. Многозадачные нейросетевые архитектуры: а) полностью общие для всех корпусов данных, б) частично общие для всех корпусов данных

Использование нескольких корпусов является многозадачным подходом, так как корпуса могут содержать неоднородные данные [110]. Проблема может встречаться при разработке многомодальных многозадачных систем в контексте одного корпуса в условиях отсутствия меток для одной из задач или модальностей [109].

Многозадачные системы могут быть и иерархическими, которые отличаются последовательным моделированием нескольких аффективных состояний, причем на каждом последующем этапе используются результаты распознавания предыдущего этапа [111]. Пример иерархической структуры для распознавания аффективных состояний показан на рисунке 7.

Преимуществом использования многозадачных и иерархических подходов к распознаванию аффективных состояний является возможность извлекать новые типы знаний, в том числе о влиянии, корреляции и взаимодействии нескольких аффективных состояний друг на друга, что потенциально влечет к улучшению качества распознавания.

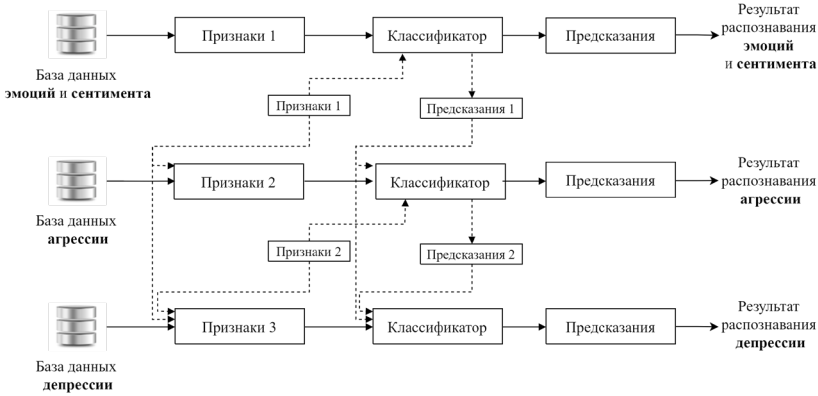

Рис. 7. Архитектура для иерархического распознавания множественных аффективных состояний

Можно сделать вывод, что область анализа эмоций и сентимента достаточно развита, существует большое количество эмоциональных и сентимент корпусов как одномодальных, так и многомодальных, а автоматические системы достигают точности около 80% распознавания эмоций и сентимента. Однако задача распознавания депрессии и агрессии значительно уступает эмоциональному распознаванию в объеме и качестве информационного и математического обеспечения. Многозадачных систем, которые распознают одновременно все четыре основных аффективных состояния (эмоции, сентимент, агрессия и депрессия), пока не существует.

5. Заключение. Исходя из аналитического обзора, можно сделать вывод, что не существует эффективных систем для одновременного распознавания различных аффективных состояний человека, поэтому ставится задача разработки автоматической программной системы, которая будет удовлетворять следующим основным требованиям:

1) Анализировать визуальную, акустическую и лингвистическую информацию (многомодальность).

2) Распознавать 4 основных аффективных состояния: эмоции, сентимент, агрессию и депрессию (многозадачность).

3) Иметь высокую точность распознавания различных видов аффектов, не менее 80% для каждого вида аффекта.

4) Работать в режиме близком к реальному времени.

5) Проводить обучение и тестирование системы на данных, полученных в реальных условиях.

Для обучения вероятностных моделей планируется использовать многомодальные корпуса CMU-MOSEI, RAMAS и DAIC, ориентированные на распознавание эмоций, сентимента и депрессии. Преимущество корпуса CMU-MOSEI заключается в том, что он имеет разметку данных на эмоции и сентимент одновременно, корпус RAMAS является единственным русскоязычным ресурсом в данной области. Корпус DAIC является самым репрезентативным информационным ресурсом по сравнению с другими многомодальными корпусами для анализа депрессии, поэтому он также будет использован для дальнейших исследований. Для решения задачи распознавания агрессии планируется собрать и аннотировать собственную базу данных.

Все это позволит разработать систему автоматического распознавания аффективных состояний человека на основе многомодального и многозадачного подхода. Аудио-, видео- и текстовые модальности оказались наиболее репрезентативными для анализа аффективных состояний. Это приводит к выводу, что данные модальности необходимо использовать в рамках многомодального подхода, предполагающего слияние модальностей на уровне прогнозов. Использование иерархического подхода для распознавания различных аффективных состояний позволит повысить эффективность работы разрабатываемой системы. Иерархия заключается в распознавании эмоций и сентимента на первом шаге, а затем полученные результаты используются для распознавания агрессии и/или депрессии. Такая система может быть использована в колл-центрах для мониторинга эмоционального состояния операторов и клиентов, в автомобилях для определения агрессивного состояния водителя, в медицинских организациях для выявления депрессивного состояния больных.

Литература

1. Picard R.W. Affective Computing for HCI // HCI (1). 1999. С. 829-833.
2. Вилюнас В.К. Эмоции // Большой психологический словарь / Под общ. ред. Б.Г. Мещерякова, В.П. Зинченко // СПб.: Прайм-ЕВРОЗНАК. 2007. С. 565-568.
3. Крафт-Эбинг Р. Учебник психиатрии. 1897. 698 с.
4. Ильин Е.П. Эмоции и чувства. Издательский дом "Питер". 2011. 782 с.
5. Тхостов А.Ш., Колымба И.Г. Эмоции и аффекты: общепсихологический и патологический аспекты // Психологический журнал. 1998. № 4. С. 41-48.
6. Ениколопов С.Н. Понятие агрессии в современной психологии // Прикладная психология. 2001. №. 1. С. 60-72.
7. Верхоляк О.В., Карпов А.А. Глава «Автоматический анализ эмоционально окрашенной речи» в монографии «Голосовой портрет ребенка с типичным и

атипичным развитием» // Е.Е. Ляксо, О.В. Фролова, С.В. Гречаный, Ю.Н. Матвеев, О.В. Верхоляк, А.А. Карпов / под ред. Е.Е. Ляксо, О.В. Фроловой // СПб. Изд-во: Издательско-полиграфическая ассоциация высших учебных заведений. 2020. С. 204.

8. Двойникова А.А., Карпов А.А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных // Информационно-управляющие системы. 2020. №. 4 (107). С. 20-30.
9. Величко А.Н., Карпов А.А. Аналитический обзор систем автоматического определения депрессии по речи. Информатика и автоматизация. №20. 2021. С. 497-529.
10. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Publishing, Arlington, VA. 2013.
11. Tzirakis P., Trigeorgis G., Nicolaou M.A., et al. End-to-end multimodal emotion recognition using deep neural networks // IEEE Journal of Selected Topics in Signal Processing. 2017. vol. 11. no. 8. pp. 1301-1309.
12. Dhall A., Goecke R., Gedeon T. Collecting large, richly annotated facial-expression databases from movies // IEEE Multimedia. 2012. vol. 19. no. 03. pp. 34-41.
13. Kossaifi J., Tzimiropoulos G., Todorovic S., et al. AFEW-VA database for valence and arousal estimation in-the-wild // Image and Vision Computing. 2017. vol. 65. pp. 23-36.
14. Kollias D., Zafeiriou S. Aff-wild2: Extending the aff-wild database for affect recognition // arXiv preprint arXiv:1811.07770. 2018.
15. Lien J.J., Kanade T., Cohn J.F., et al. Automated facial expression recognition based on FACS action units // Proceedings of third IEEE international conference on automatic face and gesture recognition. IEEE. 1998. pp. 390-395.
16. Busso C., Bulut M., Lee C.-C., et al. IEMOCAP: Interactive emotional dyadic motion capture database // Language Resources and Evaluation. 2008. vol. 42. no. 4. pp. 335-359.
17. Ringeval F., Sonderegger A., Sauer J., et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions // Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE. 2013. pp. 1-8.
18. Kossaifi J., Walecki R., Panagakis Y., et al. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild // IEEE Transactions on Pattern Analysis & Machine Intelligence. 2021. vol. 43. no. 03. pp. 1022-1040.
19. McKeown G., Valstar M.F., Cowie R., et al. The SEMAINE corpus of emotionally coloured character interactions // Proceedings of the IEEE International Conference on Multimedia and Expo. IEEE. 2010. pp. 1079-1084.
20. Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing // Proceedings of the International Conference on Speech and Computer. Springer, Cham. 2018. pp. 501-510.
21. Poria S., Hazarika D., Majumder N., et al. Meld: A multimodal multi-party dataset for emotion recognition in conversations // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. pp. 527-536.
22. Zadeh A.B., Liang P.P., Poria S., et al. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. pp. 2236-2246.
23. Pérez-Rosas V., Mihalcea R., Morency L.P. Utterance-level multimodal sentiment analysis // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. vol. 1. 2013. pp. 973-982.

24. Zadeh A., Zellers R., Pincus E., et al. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages // IEEE Intelligent Systems. 2016. vol. 31. no. 6. pp. 82-88.
25. Morency L.P., Mihalcea R., Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web // Proceedings of the 13th International Conference on Multimodal Interfaces. 2011. pp. 169-176.
26. Yu W., Xu H., Meng F., et al. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 3718-3727.
27. Lefter I., Rothkrantz L.J.M., Burghouts G., et al. Addressing multimodality in overt aggression detection // Proceedings of the International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg. 2011. pp. 25-32.
28. Lefter I., Burghouts G.J., Rothkrantz L.J.M. An audio-visual dataset of human–human interactions in stressful situations // Journal on Multimodal User Interfaces. 2014. vol. 8. no. 1. pp. 29-41.
29. Lefter I., Rothkrantz L.J.M. Multimodal cross-context recognition of negative interactions // Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE. 2017. pp. 56-61.
30. Lefter I., Jomker C.M., Tuente S.K., et al. NAA: A multimodal database of negative affect and aggression // Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE. 2017. pp. 21-27.
31. Nievas E.B., Déniz-Suárez O., Garcia G., et al. Violence detection in video using computer vision techniques // Proceedings of the International conference on Computer analysis of images and patterns. Springer, Berlin, Heidelberg. 2011. pp. 332-339.
32. Perez M., Kot A.C., Rocha A. Detection of real-world fights in surveillance videos // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2019. pp. 2662-2666.
33. Cheng M., Cai K., Li M. Rwf-2000: An open large scale video database for violence detection // Proceedings of the 25th International Conference on Pattern Recognition (ICPR). IEEE. 2021. pp. 4183-4190.
34. Kumar R., Reganti A.N., Bhatia A., et al. Aggression-annotated corpus of hindi-english code-mixed data // arXiv preprint arXiv:1803.09402. 2018.
35. Bozyiğit A., Utku S., Nasibov E. Cyberbullying detection: Utilizing social media features // Expert Systems with Applications. 2021. vol. 179. p. 115001.
36. Gratch J., Artstein R., Lucas G., et al. The Distress Analysis Interview Corpus of Human and Computer Interviews // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland. 2014. pp. 3123-3128.
37. Valstar M., Schuller B., Smith K., et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge // Proceedings of the 3rd ACMinternational workshop on Audio/visual emotion challenge (AVEC'13). Association for Computing Machinery, New York, NY, USA. 2013. pp. 3–10.
38. Yang Y., Fairbairn C., Cohn J. Detecting depression severity from vocal prosody // IEEE Transactions on Affective computing. 2013. vol. 4. no. 2. pp. 142–150.
39. Alghowinem S., Goecke R., Wagner M., et al. From joyous to clinically depressed: Mood detection using spontaneous speech // Proceedings of FLAIRS Conference, G.M. Youngblood and P.M. McCarthy, Eds. AAAI Press. 2012. pp. 141–146.
40. Huang Z., Epps J., Joachim D., et al. Depression detection from short utterances via diverse smartphones in natural environmental conditions // Proceedings of Interspeech. 2018. pp. 3393–3397.

41. Ryumina E., Karpov A. Facial expression recognition using distance importance scores between facial landmarks // CEUR Workshop Proceedings. 2020, vol. 274. pp. 1-10.
42. Axyonov, A., Ryumin, D., Kagirov, I. Method of Multi-Modal Video Analysis of Hand Movements for Automatic Recognition of Isolated Signs of Russian Sign Language // Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. 2021. vol. XLIV-2/W1-2021. pp. 7–13.
43. He K., Zhang X., Ren S., et al. Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 770–778.
44. Simonyan K., Zisserman A. Very deep convolutional networks for largescale image recognition // Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2015. pp. 1–14.
45. Niu B., Gao Z., Guo B. Facial expression recognition with LBP and ORB features // Computational Intelligence and Neuroscience. 2021. vol. 2021.
46. Verma S., Wang J., Ge Zh., et al. Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis // Proceedings of IEEE International Conference on Data Mining (ICDM). IEEE. 2020. pp. 561-570.
47. Eyben F., Weninger, F., Gross, F., et al. Recent developments in opensmile, the munich open-source multimedia feature extractor // Proceedings of ACM International Conference on Multimedia. 2013. pp. 835–838.
48. Schuller B.W., Batliner A., Bergler C., et al. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates // Proceedings of Interspeech. 2021. pp. 431–435.
49. Eyben F., Scherer K.R., Schuller, B.W., et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing // IEEE transactions on affective computing. 2015. vol. 7. no. 2. pp. 190–202.
50. Schmitt M., Ringeval F., Schuller B.W. At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech // Proceedings of Interspeech. 2016. pp. 495–499.
51. Kaya H., Karpov A.A., Salah A.A. Fisher vectors with cascaded normalization for paralinguistic analysis // Proceedings of Interspeech. 2015. pp. 909–913.
52. Zhao Z., Zhao Y., Bao Z., et al. Deep spectrum feature representations for speech emotion recognition // Proceedings of Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data. 2018. pp. 27–33.
53. Freitag M., Amiriparian S., Pugachevskiy S., et al. AuDeep: Unsupervised learning of representations from audio with deep recurrent neural networks // The Journal of Machine Learning Research. 2017. vol. 18. no. 1. pp. 6340–6344.
54. Shor J., Jansen A., Maor R., et al. Towards Learning a Universal Non-Semantic Representation of Speech // Proceedings of Interspeech. 2020. pp. 140–144.
55. Wagner J., Triantafyllopoulos A., Wierstorf H., et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap // arXiv preprint arXiv:2203.07378. 2022. pp. 1-25.
56. Degottex G., Kane J., Drugman T., et al. COVAREP – A collaborative voice analysis repository for speech technologies // Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 960-964.
57. Sogancioglu G., Verkholyak O., Kaya H., et al. Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition // Proceedings of Interspeech. 2020. pp. 2097-2101.
58. Sebastian J., Pierucci P. Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts // Proceedings of Interspeech. 2019. pp. 51-55.

59. Xu H., Zhang H., Han K., et al. Learning alignment for multimodal emotion recognition from speech // arXiv preprint arXiv:1909.05645. 2019.
60. Dellaert F., Polzin T., Waibel A. Recognizing emotion in speech // Proceedings of the 4th Int. Conf. Spoken Lang. Process (ICSLP). 1996. pp. 1970–1973.
61. Neiberg D., Elenius K., Laskowski K. Emotion recognition in spontaneous speech using GMMs // Proceedings of the 9th Int. Conf. Spoken Lang. Process. 2006. pp. 809–812.
62. Nogueiras A., Moreno A., Bonafonte A., et al. Speech emotion recognition using hidden Markov models // Proceedings of the 7th Eur. Conf. Speech Commun. Technol. 2001. pp. 746–749.
63. Raudys Š. On the universality of the single-layer perceptron model // Neural Networks and Soft Computing. Physica, Heidelberg. 2003. pp. 79-86.
64. Wang J., Lu S., Wang S.-H., et al. A review on extreme learning machine // Multimedia Tools and Applications. 2021. pp. 1-50.
65. Kruse R., Borgelt C., Klawonn F., et al. Multi-layer perceptrons // Computational Intelligence. Springer, Cham. 2022. pp. 53-124.
66. Sainath T.N., Vinyals O., Senior A., et al. Convolutional, long short-term memory, fully connected deep neural networks // Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4580–4584.
67. Kim J., Truong K.P., Englebienne G., et al. Learning spectro-temporal features with 3D CNNs for speech emotion recognition // Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII). 2017. pp. 383–388.
68. Chao L., Tao J., Yang M., et al. Long short term memory recurrent neural network based multimodal dimensional emotion recognition // Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. 2015. pp. 65-72.
69. Wang J., Xue M., Culhane R., et al. Speech emotion recognition with dual-sequence lstm architecture // Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 6474–6478.
70. Chen Q., Huang, G. A novel dual attention-based blstm with hybrid features in speech emotion recognition // Engineering Applications of Artificial Intelligence. 2021. vol. 102. p. 104277.
71. Zhao J., Mao X., Chen L. Speech emotion recognition using deep 1d & 2d cnn lstm networks // Biomedical Signal Processing and Control. 2019. vol. 47. pp. 312–323.
72. Milner R., Jalal M.A., Ng R.W., et al. A cross-corpus study on speech emotion recognition // Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2019. pp. 304-311.
73. Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need // In Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017. vol. 30. pp. 1-11.
74. Ho N.H., Yang H.J., Kim S.H., et al. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network // IEEE Access. 2020. vol. 8. pp.61672-61686.
75. Hsu W.N., Bolte B., Tsai Y.-H.H., et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. vol. 29. pp. 3451–3460.
76. Siriwardhana S., Reis A., Weerasekera R., et al. Jointly fine-tuning “bert-like” self-supervised models to improve multimodal speech emotion recognition // arXiv preprint arXiv:2008.06682. 2020. pp. 1-5.
77. Kratzwald B., Ilic S., Kraus M., et al. Deep learning for affective computing: Text-based emotion recognition in decision support // Decision Support Systems. 2018. vol. 115. pp. 24–35.

78. Stappen L., Baird A., Christ L., et al. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress // Proceedings of the 29th ACM International Conference on Multimedia (ACM MM). 2021. pp. 5706–5707.
79. Dresvyanskiy D., Ryumina E., Kaya H., et al. 2022. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild // Multimodal Technologies and Interaction. vol 6. no. 2. pp. 11.
80. Fedotov D., Kaya H., Karpov A. Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup // Proceedings of 20th International Conference on Speech and Computer (SPECOM-2018). 2018. pp. 155-165.
81. Wu C.H., Lin J.C., Wei W.L. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies // APSIPA Transactions on Signal and Information Processing. 2014. vol. 3. pp. 18.
82. Al Osman H., Falk T.H. Multimodal affect recognition: Current approaches and challenges // Emotion and Attention Recognition Based on Biological Signals and Images. 2017. pp. 59-86.
83. Liu D., Wang Z., Wang L., et al. Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning // Frontiers in Neurorobotics. 2021. pp. 13.
84. Zhang C., Yang Z., He X., et al. Multimodal intelligence: Representation learning, information fusion, and applications // IEEE Journal of Selected Topics in Signal Processing. 2020. vol. 14. no. 3. pp. 478-493.
85. Markitantov M., Ryumina E., Ryumin D., et al. Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS) Corpus: Multimodal Mask Type Recognition Task // Proceedings of Interspeech. 2022. pp. 1756-1760.
86. Yang L., Sahli H., Xia X., et al. Hybrid depression classification and estimation from audio video and text information // Proceedings of 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC). 2017. pp. 45-51.
87. Mai S., Hu H., Xing S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion // Proceedings of the AAAI Conference on Artificial Intelligence. 2020. vol. 34. no. 01. pp. 164-172.
88. Ghosal D., Akhtar M.S., Chauhan D., et al. Contextual inter-modal attention for multi-modal sentiment analysis // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2018. pp. 3454-3466.
89. Akhtar M.S., Chauhan D.S., Ghosal D., et al. Multi-task learning for multi-modal emotion recognition and sentiment analysis // arXiv preprint arXiv:1905.05812. 2019. pp. 1-10.
90. Sun Z., Sarma P, Sethares W., et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis // Proceedings of the AAAI Conference on Artificial Intelligence. 2020. vol. 34. no. 05. pp. 8992-8999.3.
91. Mai S., Hu H., Xing S. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. pp. 481-492.
92. Chauhan D.S., Akhtar M.S., Ekbal A., et al. Context-aware interactive attention for multi-modal sentiment and emotion analysis // Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. pp. 5647-5657.

93. Delbrouck J.B., Tits N., Brousmiche M., et al. A transformer-based joint-encoding for emotion recognition and sentiment analysis // arXiv preprint arXiv:2006.15955. 2020.
94. Khare A., Parthasarathy S., Sundaram S. Self-Supervised learning with cross-modal transformers for emotion recognition // IEEE Spoken Language Technology Workshop (SLT). IEEE. 2021. pp. 381-388.
95. Zaib S., Asif M., Arooj M. Development of Aggression Detection Technique in Social Media // International Journal of Information Technology and Computer Science. 2019. vol. 5. no. 8. pp. 40-46.
96. Левоневский Д.К., Савельев А.И. Подход и архитектура для систематизации и выявления признаков агрессии в русскоязычном текстовом контенте // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2021. № 54. С. 56-64.
97. Sadiq S., Mehmood A., Ullah S., et al. Aggression detection through deep neural model on twitter // Future Generation Computer Systems. 2021. vol. 114. pp. 120-129.
98. Tommasel A., Rodriguez J.M., Godoy D. Textual Aggression Detection through Deep Learning // TRAC@ COLING 2018. 2018. pp. 177-187.
99. Mandl T., Modha S., Shahi G.K., et al. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages // arXiv preprint arXiv:2112.09301. 2021.
100. Potharaju Y., Kamsali M., Kesavari C.R. Classification of Ontological Violence Content Detection through Audio Features and Supervised Learning // International Journal of Intelligent Engineering and Systems. 2019. vol. 12. no. 3. pp. 20-30.
101. Sahoo S., Routray A. Detecting aggression in voice using inverse filtered speech features // IEEE Transactions on Affective Computing. 2016. vol. 9. no. 2. pp. 217-226.
102. Santos F., Durães D., Marcondes F.M., et al. In-car violence detection based on the audio signal // Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning. Springer, Cham. 2021. pp. 437-445.
103. Liang Q., Li Y., Chen B., et al. Violence behavior recognition of two-cascade temporal shift module with attention mechanism // Journal of Electronic Imaging. 2021. vol. 30. no. 4. pp. 043009.
104. Уздяев М.Ю. Нейросетевая модель многомодального распознавания человеческой агрессии // Вестник КРАУНЦ. Физико-математические науки. 2020. Т. 33. №. 4. С. 132-149.
105. Yao Y., Papakostas M., Burzo M., et al. MUSER: MUltimodal Stress detection using Emotion Recognition as an Auxiliary Task // Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2021. pp. 2714-2725.
106. Sangwan S., Chauhan D.S., Akhtar M., et al. December. Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis // Proceedings of International Conference on Neural Information Processing. 2019. pp. 662-669.
107. Kollias D., Zafeiriou S. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface // arXiv preprint arXiv:1910.04855. 2019. pp. 1-15.
108. Li Y., Zhao T., Kawahara T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning // Proceedings of Interspeech. 2019. pp. 2803-2807.
109. Yu W., Xu H., Yuan Z., et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis // Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35. no. 12. pp. 10790-10797.
110. Vu M.T., Beurton-Aimar M., Marchand S. Multitask multi-database emotion recognition // Proceedings of IEEE/CVF International Conference on Computer Vision. 2021. pp. 3637-3644.

111. Velichko A., Markitantov M., Kaya H., et al. Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework // Proceedings of Interspeech. 2022. pp. 4735-4739.

Двойникова Анастасия Александровна — младший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: искусственный интеллект, машинное обучение, нейронные сети, сентимент-анализ, анализ аффективных состояний человека. Число научных публикаций — 10. dvoynikova.a@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Маркитантов Максим Викторович — младший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: искусственный интеллект, машинное обучение, речевые технологии, компьютерная паралингвистика, распознавание характеристик диктора, распознавание пола и возраста диктора, обнаружение защитных масок по аудиоинформации. Число научных публикаций — 11. m.markitantov@yandex.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Рюмина Елена Витальевна — младший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: аффективные вычисления, цифровая обработка изображений, распознавание аудиовизуальных сигналов, распознавание паралингвистических явлений, распознавание визуальной речи, машинное обучение, нейронные сети, биометрические системы, человеко-машинные интерфейсы. Число научных публикаций — 15. ryumina.e@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Уздяев Михаил Юрьевич — младший научный сотрудник, лаборатория технологий больших данных социокиберфизических систем, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: многомодальный анализ поведенческой активности пользователей, цифровая обработка изображений и видеосигнала, машинное обучение, нейронные сети, биометрические системы. Число научных публикаций — 20. uzdyaev.m@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Величко Алёна Николаевна — младший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: машинное обучение, речевые технологии, компьютерная паралингвистика, определение деструктивных паралингвистических явлений в разговорной речи человека, определение депрессии по разговорной речи человека. Число научных публикаций — 12. velichko.a.n@mail.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Рюмин Дмитрий Александрович — канд. техн. наук, старший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: цифровая обработка изображений, распознавание образов,

автоматическое распознавание визуальной речи, многомодальные интерфейсы, машинное обучение, нейронные сети, биометрия, человеко-машинные интерфейсы. Число научных публикаций — 62. ryumin.d@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Ляксо Елена Евгеньевна — д-р биол. наук, профессор, профессор, кафедра высшей нервной деятельности и психофизиологии биологического ф-та, Санкт-Петербургский государственный университет; ведущий научный сотрудник, лаборатория речевых и многомодальных интерфейсов, СПб ФИЦ РАН. Область научных интересов: акустика речи, детская речь, психофизиология речи. Число научных публикаций — 298. lyakso@gmail.com; Университетская наб., 7-9, 199034, Санкт-Петербург, Россия; р.т.: +7(921)996-24-92.

Карпов Алексей Анатольевич — д-р техн. наук, профессор, руководитель лаборатории, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: речевые технологии, автоматическое распознавание речи, обработка аудиовизуальной речи, многомодальные человеко-машинные интерфейсы, компьютерная паралингвистика и другие. Число научных публикаций — 350. karpov@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Поддержка исследований. Данное исследование выполнено при поддержке Российского научного фонда, проект № 22-11-00321.

DOI 10.15622/ia.21.6.2

A. DVOYNIKOVA, M. MARKITANTOV, E. RYUMINA, M. UZDIAEV,
A. VELICHKO, D. RYUMIN, E. LYAKSO, A. KARPOV

ANALYSIS OF INFOWARE AND SOFTWARE FOR HUMAN AFFECTIVE STATES RECOGNITION

Dvoynikova A., Markitantov M., Ryumina E., Uzdiaev M., Velichko A., Ryumin D., Lyakso E., Karpov A. **Analysis of infoware and software for human affective states recognition.**

Abstract. The article presents an analytical review of research in the affective computing field. This research direction is a component of artificial intelligence, and it studies methods, algorithms and systems for analyzing human affective states during interactions with other people, computer systems or robots. In the field of data mining, the definition of affect means the manifestation of psychological reactions to an exciting event, which can occur both in the short and long term, and also have different intensity. The affects in this field are divided into 4 types: affective emotions, basic emotions, sentiment and affective disorders. The manifestation of affective states is reflected in verbal data and non-verbal characteristics of behavior: acoustic and linguistic characteristics of speech, facial expressions, gestures and postures of a person. The review provides a comparative analysis of the existing infoware for automatic recognition of a person's affective states on the example of emotions, sentiment, aggression and depression. The few Russian-language, affective databases are still significantly inferior in volume and quality compared to electronic resources in other world languages. Thus, there is a need to consider a wide range of additional approaches, methods and algorithms used in a limited amount of training and testing data, and set the task of developing new approaches to data augmentation, transferring model learning and adapting foreign-language resources. The article describes the methods of analyzing unimodal visual, acoustic and linguistic information, as well as multimodal approaches for the affective states recognition. A multimodal approach to the automatic affective states analysis makes it possible to increase the accuracy of recognition of the phenomena compared to single-modal solutions. The review notes the trend of modern research that neural network methods are gradually replacing classical deterministic methods through better quality of state recognition and fast processing of large amount of data. The article discusses the methods for affective states analysis. The advantage of multitasking hierarchical approaches is the ability to extract new types of knowledge, including the influence, correlation and interaction of several affective states on each other, which potentially leads to improved recognition quality. The potential requirements for the developed systems for affective states analysis and the main directions of further research are given.

Keywords: affective states, affective computing, emotions, sentiment, depression, aggression, databases, computer systems.

References

1. Picard R.W. Affective Computing for HCI. HCI (1). 1999. pp. 829-833.
2. Viliūnas V.K. Jemocii. Bol'shoj psihologicheskij slovar' [Emotions. Big psychological dictionary]. SPb.: Prime Eurosign. 2007. pp. 565-568. (In Russ.)
3. Krafft-Ebing R. Lehrbuch der Psychiatrie. Stuttgart: Ferdinand Enke. 1897. 698 p.
4. Il'in E.P. Jemocii i chuvstva [Emotions and feelings]. Piter. 2011. 782 p. (In Russ.).
5. Thostov A.Sh., Kolymba I.G. Jemocii i affekty: obshhepsihologicheskij i patologicheskij aspekty [Emotions and affects: general psychological and pathological aspects]. Psihologicheskij zhurnal – Psychological journal. 1998. no 4. pp. 41-48. (In Russ.).

6. Enikolopov S.N. Ponjatie agressii v sovremennoj psihologii [The concept of aggression in modern psychology]. Prikladnaja psihologija – Applied psychology. 2001. no. 1. pp. 60-72. (In Russ.).
7. Verkholyak O.V., Karpov A.A. Glava "Avtomaticheskij analiz jemocional'no okrashennoj rechi" v monografii "Golosovoj portret rebenka s tipichnym i atipichnym razvitiem" [Chapter "Automatic analysis of emotionally coloured speech" in monography "Voice portrait of a child with typical and atypical development"]. E.E. Ljakso, O.V. Frolova, S.V. Grechanyj, Ju.N. Matveev, O.V. Verkholyak, A.A. Karpov. SPb. Izdatel'sko-poligraficheskaja associacija vysshih uchebnyh zavedenij [Publishing and Printing Association of Higher Educational Institutions]. 2020. 204 p. (In Russ.).
8. Dvoynikova A. A., Karpov A. A. Analytical review of approaches to Russian text sentiment recognition. Informatsionno-upravliaiushchie sistemy – Information and Control Systems. 2020. no. 4. pp. 20–30. (In Russ.).
9. Velichko A., Karpov A. Analytical Review of Automatic Systems for Depression Detection by Speech. Informatics and Automation. 2021. no. 20. pp. 497-529. (In Russ.).
10. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Publishing, Arlington, VA. 2013.
11. Tzirakis P., Trigeorgis G., Nicolaou M.A., et al. End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing. 2017. vol. 11. no. 8. pp. 1301-1309.
12. Dhall A., Goecke R., Gedeon T. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia. 2012. vol. 19. no. 03. pp. 34-41.
13. Kossaifi J., Tzimiropoulos G., Todorovic S., et al. AFEW-VA database for valence and arousal estimation in-the-wild. Image and Vision Computing. 2017. vol. 65. pp. 23-36.
14. Kollias D., Zafeiriou S. Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770. 2018.
15. Lien J.J., Kanade T., Cohn J.F., et al. Automated facial expression recognition based on FACS action units. Proceedings of third IEEE international conference on automatic face and gesture recognition. IEEE. 1998. pp. 390-395.
16. Busso C., Bulut M., Lee C.-C., et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 2008. vol. 42. no. 4. pp. 335-359.
17. Ringeval F., Sonderegger A., Sauer J., et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE. 2013. pp. 1-8.
18. Kossaifi J., Walecki R., Panagakis Y., et al. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2021. vol. 43. no. 03. pp. 1022-1040.
19. McKeown G., Valstar M.F., Cowie R., et al. The SEMAINE corpus of emotionally coloured character interactions. Proceedings of the IEEE International Conference on Multimedia and Expo. IEEE. 2010. pp. 1079-1084.
20. Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. Proceedings of the International Conference on Speech and Computer. Springer, Cham. 2018. pp. 501-510.
21. Poria S., Hazarika D., Majumder N., et al. Meld: A multimodal multi-party dataset for emotion recognition in conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. pp. 527-536.

22. Zadeh A.B., Liang P.P., Poria S., et al. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. pp. 2236-2246
23. Pérez-Rosas V., Mihalcea R., Morency L.P. Utterance-level multimodal sentiment analysis. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. vol. 1. 2013. pp. 973-982.
24. Zadeh A., Zellers R., Pincus E., et al. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. EEE Intelligent Systems. 2016. vol. 31. no. 6. pp. 82-88.
25. Morency L.P., Mihalcea R., Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web. Proceedings of the 13th International Conference on Multimodal Interfaces. 2011. pp. 169-176.
26. Yu W., Xu H., Meng F., et al. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 3718-3727.
27. Lefter I., Rothkrantz L.J.M., Burghouts G., et al. Addressing multimodality in overt aggression detection. Proceedings of the International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg. 2011. pp. 25-32.
28. Lefter I., Burghouts G.J., Rothkrantz L.J.M. An audio-visual dataset of human–human interactions in stressful situations. Journal on Multimodal User Interfaces. 2014. vol. 8. no. 1. pp. 29-41.
29. Lefter I., Rothkrantz L.J.M. Multimodal cross-context recognition of negative interactions. Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE. 2017. pp. 56-61.
30. Lefter I., Jomker C.M., Tuente S.K., et al. NAA: A multimodal database of negative affect and aggression. Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE. 2017. pp. 21-27.
31. Nievas E.B., Déniz-Suárez O., Garcia G., et al. Violence detection in video using computer vision techniques. Proceedings of the International conference on Computer Analysis of Images and Patterns. Springer, Berlin, Heidelberg. 2011. pp. 332-339.
32. Perez M., Kot A.C., Rocha A. Detection of real-world fights in surveillance videos. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2019. pp. 2662-2666.
33. Cheng M., Cai K., Li M. Rwf-2000: An open large scale video database for violence detection. Proceedings of the 25th International Conference on Pattern Recognition (ICPR). IEEE. 2021. pp. 4183-4190.
34. Kumar R., Reganti A.N., Bhatia A., et al. Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402. 2018.
35. Bozyiğit A., Utku S., Nasibov E. Cyberbullying detection: Utilizing social media features. Expert Systems with Applications. 2021. vol. 179. p. 115001.
36. Gratch J., Artstein R., Lucas G., et al. The Distress Analysis Interview Corpus of Human and Computer Interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland. 2014. pp. 3123-3128.
37. Valstar M., Schuller B., Smith K., et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACMinternational workshop on Audio/visual emotion challenge (AVEC'13). Association for Computing Machinery, New York, NY, USA. 2013. pp. 3–10.
38. Yang Y., Fairbairn C., Cohn J. Detecting depression severity from vocal prosody. IEEE Transactions on Affective computing. 2013. vol. 4. no. 2. pp. 142–150.

39. Alghowinem S., Goecke R., Wagner M., et al. From joyous to clinically depressed: Mood detection using spontaneous speech. Proceedings of FLAIRS Conference, G.M. Youngblood and P.M. McCarthy, Eds. AAAI Press. 2012. pp. 141–146.
40. Huang Z., Epps J., Joachim D., et al. Depression detection from short utterances via diverse smartphones in natural environmental conditions. Proceedings of Interspeech. 2018. pp. 3393–3397.
41. Ryumina E., Karpov A. Facial expression recognition using distance importance scores between facial landmarks. CEUR Workshop Proceedings. 2020, vol. 2744. pp. 1-10.
42. Axyonov, A., Ryumin, D., Kagirov, I. Method of Multi-Modal Video Analysis of Hand Movements for Automatic Recognition of Isolated Signs of Russian Sign Language. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. 2021. vol. XLIV-2/W1-2021. pp. 7–13.
43. He K., Zhang X., Ren S., et al. Deep residual learning for image recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 770–778.
44. Simonyan K., Zisserman A. Very deep convolutional networks for largescale image recognition. Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2015. pp. 1–14.
45. Niu B., Gao Z., Guo B. Facial expression recognition with LBP and ORB features. Computational Intelligence and Neuroscience. 2021. vol. 2021.
46. Verma S., Wang J., Ge Zh., et al. Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis. Proceedings of IEEE International Conference on Data Mining (ICDM). IEEE. 2020. pp. 561-570.
47. Eyben F., Weninger, F., Gross, F., et al. Recent developments in opensmile, the munich open-source multimedia feature extractor. Proceedings of ACM International Conference on Multimedia. 2013. pp. 835–838.
48. Schuller B.W., Batliner A., Bergler C., et al. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. Proceedings of Interspeech. 2021. pp. 431–435.
49. Eyben F., Scherer K.R., Schuller, B.W., et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing. 2015. vol. 7. no. 2. pp. 190–202.
50. Schmitt M., Ringeval F., Schuller B.W. At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech. Proceedings of Interspeech. 2016. pp. 495–499.
51. Kaya H., Karpov A.A., Salah A.A. Fisher vectors with cascaded normalization for paralinguistic analysis. Proceedings of Interspeech. 2015. pp. 909–913.
52. Zhao Z., Zhao Y., Bao Z., et al. Deep spectrum feature representations for speech emotion recognition. Proceedings of Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data. 2018. pp. 27–33.
53. Freitag M., Amiriparian S., Pugachevskiy S., et al. AuDeep: Unsupervised learning of representations from audio with deep recurrent neural networks. The Journal of Machine Learning Research. 2017. vol. 18. no. 1. pp. 6340–6344.
54. Shor J., Jansen A., Maor R., et al. Towards Learning a Universal Non-Semantic Representation of Speech. Proceedings of Interspeech. 2020. pp. 140–144.
55. Wagner J., Triantafyllopoulos A., Wierstorf H., et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap. arXiv preprint arXiv:2203.07378. 2022. pp. 1-25.

56. Degottex G., Kane J., Drugman T., et al. COVAREP – A collaborative voice analysis repository for speech technologies. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 960-964.
57. Sogancioglu G., Verkholyak O., Kaya H., et al. Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition. Proceedings of Interspeech. 2020. pp. 2097-2101.
58. Sebastian J., Pierucci P. Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts. Proceedings of Interspeech. 2019. pp. 51-55.
59. Xu H., Zhang H., Han K., et al. Learning alignment for multimodal emotion recognition from speech. arXiv preprint arXiv:1909.05645. 2019.
60. Dellaert F., Polzin T., Waibel A. Recognizing emotion in speech. Proceedings of the 4th Int. Conf. Spoken Lang. Process (ICSLP). 1996. pp. 1970–1973.
61. Neiberg D., Elenius K., Laskowski K. Emotion recognition in spontaneous speech using GMMs. Proceedings of the 9th Int. Conf. Spoken Lang. Process. 2006. pp. 809–812.
62. Nogueiras A., Moreno A., Bonafonte A., et al. Speech emotion recognition using hidden Markov models. Proceedings of the 7th Eur. Conf. Speech Commun. Technol. 2001. pp. 746–749.
63. Raudys Š. On the universality of the single-layer perceptron model. Neural Networks and Soft Computing. Physica. Heidelberg. 2003. pp. 79-86.
64. Wang J., Lu S., Wang S.-H., et al. A review on extreme learning machine. Multimedia Tools and Applications. 2021. pp. 1-50.
65. Kruse R., Borgelt C., Klawonn F., et al. Multi-layer perceptrons. Computational Intelligence. Springer, Cham. 2022. pp. 53-124.
66. Sainath T.N., Vinyals O., Senior A., et al. Convolutional, long short-term memory, fully connected deep neural networks. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4580–4584.
67. Kim J., Truong K.P., Englebienne G., et al. Learning spectro-temporal features with 3D CNNs for speech emotion recognition. Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII). 2017. pp. 383–388.
68. Chao L., Tao J., Yang M., et al. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. 2015. pp. 65-72.
69. Wang J., Xue M., Culhane R., et al. Speech emotion recognition with dual-sequence lstm architecture. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 6474–6478.
70. Chen Q., Huang, G. A novel dual attention-based blstm with hybrid features in speech emotion recognition. Engineering Applications of Artificial Intelligence. 2021. vol. 102. p. 104277.
71. Zhao J., Mao X., Chen L. Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomedical Signal Processing and Control. 2019. vol. 47. pp. 312–323.
72. Milner R., Jalal M.A., Ng R.W., et al. A cross-corpus study on speech emotion recognition. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2019. pp. 304-311.
73. Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017. vol. 30. pp. 1-11.
74. Ho N.H., Yang H.J., Kim S.H., et al. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access. 2020. vol. 8. pp.61672-61686.

75. Hsu W.N., Bolte B., Tsai Y.-H.H., et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. vol. 29. pp. 3451–3460.
76. Siriwardhana S., Reis A., Weerasekera R., et al. Jointly fine-tuning "bert-like" self-supervised models to improve multimodal speech emotion recognition. arXiv preprint arXiv:2008.06682. 2020. pp. 1-5.
77. Kratzwald B., Ilic S., Kraus M., et al. Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems. 2018. vol. 115. pp. 24–35.
78. Stappen L., Baird A., Christ L., et al. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. Proceedings of the 29th ACM International Conference on Multimedia (ACM MM). 2021. pp. 5706–5707.
79. Dresvyanskiy D., Ryumina E., Kaya H., et al. 2022. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild. Multimodal Technologies and Interaction. vol 6. no. 2. pp. 11.
80. Fedotov D., Kaya H., Karpov A. Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup. Proceedings of the 20th International Conference on Speech and Computer (SPECOM-2018). 2018. pp. 155-165.
81. Wu C.H., Lin J.C., Wei W.L. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA Transactions on Signal and Information Processing. 2014. vol. 3. pp. 18.
82. Al Osman H., Falk T.H. Multimodal affect recognition: Current approaches and challenges. Emotion and Attention Recognition Based on Biological Signals and Images. 2017. pp. 59-86.
83. Liu D., Wang Z., Wang L., et al. Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning. Frontiers in Neurorobotics. 2021. pp. 13.
84. Zhang C., Yang Z., He X., et al. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing. 2020. vol. 14. no. 3. pp. 478-493.
85. Markitantov M., Ryumina E., Ryumin D., et al. Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS) Corpus: Multimodal Mask Type Recognition Task. Proceedings of Interspeech. 2022. pp. 1756-176.
86. Yang L., Sahli H., Xia X., et al. Hybrid depression classification and estimation from audio video and text information. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC). 2017. pp. 45-51.
87. Mai S., Hu H., Xing S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. Proceedings of the AAAI Conference on Artificial Intelligence. 2020. vol. 34. no. 01. pp. 164-172.
88. Ghosal D., Akhtar M.S., Chauhan D., et al. Contextual inter-modal attention for multi-modal sentiment analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2018. pp. 3454-3466.
89. Akhtar M.S., Chauhan D.S., Ghosal D., et al. Multi-task learning for multi-modal emotion recognition and sentiment analysis. arXiv preprint arXiv:1905.05812. 2019. pp. 1-10.
90. Sun Z., Sarma P, Sethares W., et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proceedings of the AAAI Conference on Artificial Intelligence. 2020. vol. 34. no. 05. pp. 8992-8999.3.
91. Mai S., Hu H., Xing S. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. pp. 481-492.
92. Chauhan D.S., Akhtar M.S., Ekbal A., et al. Context-aware interactive attention for multi-modal sentiment and emotion analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. pp. 5647-5657.
93. Delbrouck J.B., Tits N., Brousmiche M., et al. A transformer-based joint-encoding for emotion recognition and sentiment analysis. arXiv preprint arXiv:2006.15955. 2020.
94. Khare A., Parthasarathy S., Sundaram S. Self-Supervised learning with cross-modal transformers for emotion recognition. IEEE Spoken Language Technology Workshop (SLT). IEEE. 2021. pp. 381-388.
95. Zaib S., Asif M., Arooj M. Development of Aggression Detection Technique in Social Media. International Journal of Information Technology and Computer Science. 2019. vol. 5. no. 8. pp. 40-46.
96. Levonevskij D.K., Savel'ev A.I. Podhod i arhitektura dlja sistematizacii i vyjavlenija priznakov agressii v russkojazychnom tekstovom kontente [Approach and architecture for systematization and identification of aggression signs in Russian-language text content.]. Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naja tehnika i informatika – Tomsk state university journal of control and computer science. Control, computer engineering and informatics. 2021. no. 54. pp. 56-64. (In Russ.).
97. Sadiq S., Mehmood A., Ullah S., et al. Aggression detection through deep neural model on twitter. Future Generation Computer Systems. 2021. vol. 114. pp. 120-129.
98. Tommasel A., Rodriguez J.M., Godoy D. Textual Aggression Detection through Deep Learning. TRAC@ COLING 2018. 2018. pp. 177-187.
99. Mandl T., Modha S., Shahi G.K., et al. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. arXiv preprint arXiv:2112.09301. 2021.
100. Potharaju Y., Kamsali M., Kesavari C.R. Classification of Ontological Violence Content Detection through Audio Features and Supervised Learning. International Journal of Intelligent Engineering and Systems. 2019. vol. 12. no. 3. pp. 20-30.
101. Sahoo S., Routray A. Detecting aggression in voice using inverse filtered speech features. IEEE Transactions on Affective Computing. 2016. vol. 9. no. 2. pp. 217-226.
102. Santos F., Durães D., Marcondes F.M., et al. In-car violence detection based on the audio signal. Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning. Springer, Cham. 2021. pp. 437-445.
103. Liang Q., Li Y., Chen B., et al. Violence behavior recognition of two-cascade temporal shift module with attention mechanism. Journal of Electronic Imaging. 2021. vol. 30. no. 4. pp. 043009.
104. Uzdyaev M.Yu. Neural network model for multimodal recognition of human aggression. Bulletin KRASEC. Physical and Mathematical Sciences. 2020. vol. 33. no. 4. pp. 132-149. (In Russ.).
105. Yao Y., Papakostas M., Burzo M., et al. MUSER: MULtimodal Stress detection using Emotion Recognition as an Auxiliary Task. Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2021. pp. 2714-2725.
106. Sangwan S., Chauhan D.S., Akhtar M., et al. December. Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis. Proceedings of International Conference on Neural Information Processing. 2019. pp. 662-669.
107. Kollias D., Zafeiriou S. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855. 2019. pp. 1-15.

108. Li Y., Zhao T., Kawahara T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. Proceedings of Interspeech. 2019. pp. 2803-2807.
109. Yu W., Xu H., Yuan Z., et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35. no. 12. pp. 10790-10797.
110. Vu M.T., Beurton-Aimar M., Marchand S. Multitask multi-database emotion recognition. Proceedings of IEEE/CVF International Conference on Computer Vision. 2021. pp. 3637-3644.
111. Velichko A., Markitantov M., Kaya H., et al. Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework. Proceedings of Interspeech. 2022. pp. 4735-4739.

Dvoynikova Anastasia — Junior researcher, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: artificial intelligence, machine learning, neural networks, recognition of protective masks by audio information, sentiment analysis, human affective states analysis. The number of publications — 10. dvoynikova.a@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Markitantov Maxim — Junior researcher, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: artificial intelligence, machine learning, speech technologies, computational paralinguistics, recognition of the speaker's characteristics, speaker's age and gender recognition, detection of protective masks by audio information. The number of publications — 11. m.markitantov@yandex.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Ryumina Elena — Junior researcher, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: affective computing, digital image processing, audio-visual signal recognition, recognition of paralinguistic phenomena, visual speech recognition, machine learning, neural networks, biometric systems, human-machine interfaces. The number of publications — 15. ryumina.e@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Uzdiaev Mikhail — Junior researcher, Laboratory of big data technologies in socio-cyberphysical systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: multimodal analysis of user behavior activity, digital image processing, digital video processing, machine learning, neural networks, biometric systems. The number of publications — 20. uzdyaev.m@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Velichko Alena — Junior researcher, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: machine learning, speech technologies, computational paralinguistics, detection of destructive paralinguistic phenomena in colloquial speech, depression detection in colloquial speech. The number of publications — 12. velichko.a.n@mail.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Ryumin Dmitry — Ph.D., Senior researcher, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS).

Research interests: digital image processing, pattern recognition, automatic visual speech recognition, multimodal interfaces, machine learning, neural networks, biometrics, human-machine interfaces. The number of publications — 62. ryumin.d@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Lyakso Elena — Ph.D., Dr.Sci., Professor, Professor, Department of higher nervous activity and psychophysiology, biology faculty, St. Petersburg State University; Leading researcher, Laboratory of speech and multimodal interfaces, SPC RAS. Research interests: speech acoustic, child speech, speech psychophysiology. The number of publications — 298. lyakso@gmail.com; 7-9, University Emb., 199034, St. Petersburg, Russia; office phone: +7(921)996-24-92.

Karpov Alexey — Ph.D., Dr.Sci., Professor, Head of laboratory, Laboratory of speech and multimodal interfaces, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: speech technology, automatic speech recognition, audio-visual speech processing, multimodal human-computer interfaces, and computational paralinguistics. The number of publications — 350. karpov@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Acknowledgements. This research was supported by Russian Science Foundation (grant № 22-11-00321).