БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УДК 004.5

Бушлякова Маргарита Дмитриевна

Современные информационные технологии распознавания эмоций в человеко-компьютерном взаимодействии

Реферат по дисциплине
«Основы информационных технологий»

Магистранта кафедры информационных систем
управления факультета прикладной
математики и информатики

Специальность: 7-06-0533-05
+375445496410
margo.bushliakova@gmail.com

Рецензент:

_____

Минск, 2025

СОДЕРЖАНИЕ

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

AER — Affective Emotion Recognition, распознавание эмоций
AWGN — Additive White Gaussian Noise, аддитивный белый гауссов шум
BERT — Bidirectional Encoder Representations from Transformers, трансформерная модель
CK+ — база данных эмоциональных выражений (Cohn–Kanade Extended)
CLIP — Contrastive Language–Image Pre-training, модель контрастивного обучения
CNN — Convolutional Neural Network, свёрточная нейронная сеть
CNN–LSTM — гибридная архитектура свёрточной и рекуррентной сетей
DT — Decision Tree, дерево решений
EfficientNet — архитектура свёрточных нейронных сетей
EI — Emotional Intelligence, эмоциональный интеллект
Emotion-Specific Attention — механизм внимания, специфичный для эмоций
FACS — Facial Action Coding System, система кодирования лицевых движений
FER — Facial Emotion Recognition, распознавание эмоций по лицу
Fine-tuned — режим дообучения модели
GFLOPs — Giga Floating Point Operations per Second, миллиард операций с плавающей запятой
HCI — Human–Computer Interaction, взаимодействие человека с компьютером
HFE-Net — Hybrid Feature Extraction Network, гибридная архитектура извлечения признаков
IEMOCAP — интерактивный корпус аудио- и видеозаписей эмоций
“in the wild” — условия съёмки в реальной среде
LBP — Local Binary Patterns, локальные двоичные шаблоны
LLaMA — Large Language Model Meta AI, крупная языковая модель
LoRA — Low-Rank Adaptation, параметро-эффективное дообучение
macro-F1 — макроусреднённая F1-мера
MER — Multimodal Emotion Recognition, мультимодальное распознавание эмоций
MER-CLIP — мультимодальная модель на основе CLIP
MLLM — Multimodal Large Language Model, мультимодальная большая языковая модель
MSP-PODCAST — база данных эмоциональной речи
NCDE — Neural Controlled Differential Equations, нейронные управляемые дифференциальные уравнения
NKF — New Kernel-based Framework, ядерная архитектура
NLP-based AER — распознавание эмоций в тексте на основе NLP
PET — Parameter-Efficient Tuning, параметро-эффективное дообучение

ResNet — архитектура Residual Network
RF — Random Forest, случайный лес
RNN — Recurrent Neural Network, рекуррентная нейронная сеть
RoBERTa — улучшенная модель на основе BERT
SALMONN — мультимодальная модель обработки аудио
SER — Speech Emotion Recognition, распознавание эмоций по речи
SSL — Self-Supervised Learning, самообучение без учителя
“state-of-the-art “— передовой уровень техники
SVM — Support Vector Machine, метод опорных векторов
TF-IDF–Gating — механизм гейтинга на основе TF–IDF
UX — User Experience, пользовательский опыт
VEGA — модель мультимодального анализа эмоций
Vector Scaling — метод калибровки моделей
Vision Transformer — архитектура трансформеров для компьютерного зрения
WA — Weighted Accuracy, взвешенная точность
Wav2Vec2.0 — модель самообучения для аудио
WavLM — модель самосупервизированного обучения речи

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Общий объем реферата составляет 28 страниц, номинальный объём 17 страниц, 2 приложения. Библиографический список включает 34 источника. Таблицы и графические материалы в работе отсутствуют.

Реферат представляет собой комплексное исследование, направленное на систематический анализ и критическую оценку современных информационных технологий в области эмоционального искусственного интеллекта. Работа последовательно рассматривает ключевые методы, архитектуры и практические решения, определяющие текущее состояние и перспективы развития информационных технологий в данной предметной области.

Ключевые слова: распознавание эмоций, человеко-компьютерное взаимодействие (HCI), эмоциональный интеллект (EI), мультимодальный анализ, распознавание эмоций по лицу (FER), распознавание эмоций по речи (SER), анализ эмоций в тексте (AER), мультимодальное распознавание эмоций (MER), нейронные сети, глубокое обучение.

Целью исследования является комплексный анализ современных информационных технологий распознавания эмоций, а также оценка их применимости, фундаментальных ограничений и перспектив развития в контексте систем человеко-компьютерного взаимодействия (HCI).

Объектом исследования выступают методы, нейросетевые архитектуры и системы, предназначенные для распознавания эмоциональных состояний человека по различным модальностям: визуальным (FER), аудио (SER), текстовым (AER) и мультимодальным (MER) данным, исследование которых выявляет ряд фундаментальных проблем, определяющих актуальность работы.

Актуальность исследования обусловлена совокупностью стратегических, научных и технологических факторов. Во-первых, эмоциональный интеллект (EI) становится ключевым фактором повышения качества человеко-компьютерного взаимодействия. Системы, способные распознавать и адаптироваться к эмоциональному состоянию пользователя, обеспечивают более высокий уровень пользовательского опыта (UX), повышают эффективность чат-ботов, образовательных платформ и цифровых сервисов. Во-вторых, наблюдается высокий научный интерес к данной области, что подтверждается «взрывным ростом» числа публикаций после 2020 года, обусловленным прогрессом в области глубоких нейросетевых архитектур. В-третьих, мультимодальность, объединяющая данные из разных источников, превратилась в ключевой технологический тренд. Согласно прогнозу Gartner, к 2027 году 40% решений генеративного ИИ будут мультимодальными, что подчёркивает стратегическую значимость исследований в этой сфере. Наконец, несмотря на достигнутые успехи, точность и надёжность распознавания эмоций

в реальных условиях остаются нерешённой проблемой, что делает исследования, направленные на преодоление существующих технологических барьеров, крайне востребованными.

Во введении обосновывается актуальность темы, определяется её теоретическая база на основе классических моделей эмоций Экмана, Плучика и Рассела, а также подчёркивается центральная роль мультимодального анализа в современных системах эмоционального ИИ.

В Главе 1 проводится исчерпывающий критический анализ, систематизирующий ключевые архитектурные подходы (FER, SER, AER, MER) и выявляющий общие технологические барьеры, такие как разрыв в производительности моделей в контролируемых и реальных условиях.

В Главе 2 описывается методологический аппарат, включающий систематический литературный анализ и сравнительную оценку нейросетевых архитектур по критериям точности, обобщающей способности и применимости в реальных условиях HCI.

В Главе 3 кристаллизуется центральный тезис работы: неконтролируемое усложнение архитектур не обеспечивает пропорционального роста качества распознавания, уступая в эффективности семантически-ориентированным и параметро-эффективным подходам.

В заключении подводятся итоги исследования, обобщаются выявленные закономерности и определяются стратегические направления для дальнейшего развития эмоционально-интеллектуальных интерфейсов.

В приложении А слайды презентации реферата.

В приложении Б персональный сайт.

ВВЕДЕНИЕ

Эмоциональный интеллект (EI) в человеко-компьютерном взаимодействии (HCI) становится ключевым фактором повышения качества и естественности коммуникации между пользователем и системой. Источники [13] подчёркивают, что интерфейсы, не учитывающие эмоциональное состояние пользователя, утрачивают способность к адаптивному и ситуативно корректному взаимодействию, что снижает эффективность UX и ограничивает потенциал цифровых сервисов. Практические исследования [34] и работы Pantic и Rothkrantz [20] демонстрируют, что эмоционально чувствительные интерфейсы способны корректировать стиль общения, сложность контента и поведенческие механики, повышая вовлечённость и удовлетворённость пользователей. В сфере чат-ботов EI напрямую связан с эффективностью поддержки: исследования [8] показывают, что эмоционально адаптивные боты снижают стресс, улучшают восприятие сервиса и ускоряют решение проблем. Аналогично, в образовательных и когнитивных системах EI способствует повышению мотивации и продуктивности взаимодействия [16].

Актуальность исследования анализа эмоций подтверждается значительным ростом числа работ за последние годы. В систематических обзорах [33], [30], [26] отмечается «взрывной рост» публикаций после 2020 года, связанный с развитием глубоких нейросетевых архитектур и внедрением мультимодальных методов. Эти обзоры подчёркивают, что эмоциональное распознавание превратилось в одно из центральных направлений HCI, компьютерного зрения, UX-аналитики и разработки диалоговых систем.

Теоретическую основу эмоционального анализа составляют классические модели классификации эмоций. Модель Экмана описывает набор универсальных эмоций, применяемый как базовый стандарт в распознавании выражений лица. Таксономия Плучика организует эмоции в виде «эмоционального колеса», позволяя представлять их интенсивность и противоположность. Модель Рассела (Circumplex) трактует эмоции как комбинацию валентности и активации, что особенно важно при анализе динамических состояний. PAD-модель (Pleasure–Arousal–Dominance) расширяет эту структуру, вводя компонент доминантности, что делает её применимой для алгоритмов оценки взаимодействия в HCI. Эти модели формируют общую терминологическую и концептуальную основу, необходимую для построения систем эмоционального ИИ.

Особую роль в современных ИТ играет мультимодальный анализ, объединяющий визуальные, голосовые, текстовые и поведенческие сигналы. Фундаментальный обзор [6] показывает, что мультимодальность позволяет компенсировать ограничения отдельных каналов и значительно повышает

устойчивость и точность распознавания эмоций. Аналитические работы [33], [30], [5] подтверждают, что интеграция нескольких модальностей является оптимальным решением для интерфейсов, работающих в реальных условиях: при шуме, неоднозначности поведения и ограниченности данных. Индустриальные прогнозы Gartner [12] подчёркивают стратегическую значимость этого направления: к 2027 году 40% решений GenAI будут мультимодальными, что делает мультимодальный эмоциональный анализ одним из ключевых технологических трендов.

Несмотря на успехи, проблема точного распознавания эмоций остаётся одной из наиболее сложных. Современные исследования указывают на ряд ограничений: вариативность выражений у разных людей, культурные различия, шум в аудио- и видеоданных, ограниченность и несбалансированность датасетов, а также слабую переносимость моделей на новые условия и пользователей. Обзоры [33], [26] подчёркивают, что даже мультимодальные системы испытывают трудности в сценариях реального времени и требуют дальнейших исследований для повышения точности и предсказуемости.

Учитывая важность эмоционального интеллекта для современного HCI, стремительный рост научных работ и технологическую значимость мультимодальных методов, выбор темы «Современные информационные технологии распознавания эмоций в человеко-компьютерном взаимодействии» является обоснованным и актуальным. Эта тема отражает как фундаментальные теоретические вопросы анализа эмоций, так и прикладные аспекты внедрения EI в адаптивные интерфейсы, чат-боты, образовательные и сервисные системы. Исследование современных технологий эмоционального ИИ позволяет не только оценить текущее состояние области, но и определить перспективы её развития, что делает работу значимой как научно, так и на практике.

ГЛАВА 1
АНАЛИТИЧЕСКИЙ ОБЗОР ЛИТЕРАТУРЫ

1.1 Технологии и методы распознавания эмоций по визуальным данным (FER)

При подготовке аналитического обзора были проанализированы современные исследования в области распознавания эмоций по лицу. Рассматриваемые источники показывают, что развитие FER связано не столько с созданием новых монолитных архитектур, сколько с интеграцией CNN, трансформеров и лицевых ориентиров, а также с усложнением стратегий извлечения и слияния признаков.

Гибридные архитектуры FER подразумевают интеграцию CNN, SVM и Transformer. Рассмотрим некоторые такие примеры, упомянутые в работах [3], [10], [23].

Работа [3, с.14, 22] демонстрирует многоуровневый ансамбль на основе EfficientNet-B0, комбинирующий transfer learning, набор бинарных классификаторов и мета-классификатор, что позволило достичь 92% точности на CK+ [3, с.14, 22]. Достоинство подхода — уменьшение переобучения за счёт декомпозиции задачи.

Работа [10, с.1–2, 7] предлагает трёхстадийный конвейер: физически-обоснованная предобработка изображения, модифицированная ResNet-18 и финальная классификация SVM [10, с.1–2, 7]. Сильной стороной является повышение устойчивости в условиях некачественного освещения; однако метод опирается на точность этапа предобработки и требует большей вычислительной способности.

HFE-Net объединяет EfficientNetV2 и многоголовое самовнимание для моделирования дальних зависимостей на лице [23, с.1–2]. Авторы отмечают, что механизм внимания компенсирует локальность CNN, но модель может быть чувствительна к фоновому шуму и вариативности позы.

Рассмотрим теперь семантически направленное извлечение признаков. Самым заметным современным трендом является использование лицевых ориентиров. Модель NKF в работе [22, с.2–3] использует 68 ориентиров и извлекает признаки строго из значимых областей лица (глаза, брови, рот) [22, с.2–3]. Такой подход повышает интерпретируемость и робастность в условиях “in-the-wild". Авторы вводят два механизма повышения устойчивости. Первый механизм это добавление шума к координатам ориентиров, что уменьшает чувствительность к ошибкам детекции [22, с.8–9]. Второй — механизм внимания, использующий один "репрезентативный" признак для усиления других [22, с.9–10].

Проведём сравнение и анализ подходов, которые используют гибриды и те, что выполняют семантический анализ, чтобы выявить влияние архитектуры на качество FER.

Сравнение результатов на RAF-DB и AffectNet показывает, что подходы с семантически направленным извлечением признаков часто превосходят гибриды CNN+Transformer. NKF достигает 93.16% на RAF-DB и 64.87% на AffectNet-8 [22, с.12–13]. HFE-Net показывает 87.29% и 58.55% соответственно [23, с.7]. Разница указывает на то, что внимание к локальным ключевым зонам лица эффективнее, чем универсальные transformer-механизмы, особенно в условиях искажений освещения и позы.

Перейдём к рассмотрению критических проблем FER технологий, ограничений и нерешённых вопросов. Несмотря на высокие показатели на бенчмарках, все источники подчёркивают существующие проблемы:

1) условия “in-the-wild" — окклюзии, большие углы поворота головы, неравномерное освещение [10, с.3], [22, с.2].

2) присутствует дисбаланс классов и культуральная предвзятость, особенно для редких эмоций [22, с.13], [23, с.7].

3) неоднозначность отдельных эмоций, например “страх–удивление” [3, с.10].

4) высокая вычислительная сложность, поскольку NKF достигает 7.13 GFLOPs, что ограничивает применение на мобильных устройствах [22, с.16].

Перейдём к анализу тенденций и перспектив развития FER. На основании рассмотренных публикаций можно выделить несколько направлений развития:

1) Оптимизация моделей под мобильные устройства [10, с.30].

2) Повышение устойчивости к окклюзиям и позовым искажениям, включая 3D-модели и улучшенные детекторы ориентиров [22, с.20].

3) Интеграция априорных знаний о лице (семантика регионов, внимание к ключевым точкам).

4) Переход к мультимодальным системам, объединяющим визуальные, аудио- и текстовые признаки [3, с.24].

1.2 Технологии распознавания эмоций по аудиосигналам (SER)

В последние годы область распознавания эмоций по речи демонстрирует устойчивый переход от использования актерских, строго контролируемых аудиозаписей к анализу «натуралистичных» данных, отражающих реальные условия коммуникации. Этот переход обусловлен потребностью в построении более прикладных систем человеко-компьютерного взаимодействия, однако сопровождается значительным усложнением задачи вследствие высокой вариативности речевых паттернов и выраженной несбалансированности

классов [25, с.1; 1, с.1]. На этом фоне особую значимость приобретает тщательный анализ архитектурных решений и методологических подходов, позволяющих эффективно работать с данными различной степени сложности. Проведём сравнительный анализ архитектур SER.

Гибридные модели, сочетающие сверточные и рекуррентные нейронные сети, исторически сформировали основу для SER-систем, работающих со спектрограммами речевых сигналов. Исследование [25] демонстрирует, что сочетание Time Distributed 2D CNN с LSTM позволяет достичь высокой точности (96.5%), тогда как использование двунаправленного LSTM с механизмом внимания повышает результат до 98.1% [25, с.11–12].

Применение аугментации (AWGN) снижает переобучение на ограниченных актёрских выборках [25, с.8–9]. При этом данные модели демонстрируют ограниченную способность к обобщению в условиях спонтанной речи, что указывает на фундаментальную зависимость CNN–LSTM от качества и однородности данных.

Использование предобученных моделей самообучения (SSL), таких как WavLM и Wav2vec2.0, стало ключевым направлением развития SER-систем на сложных «натуралистичных» данных [1, с.2; 29, с.4].

Система "Abhinaya" [1] иллюстрирует современную тенденцию к построению ансамблей, включающих аудиомодели на базе WavLM-Large и SALMONN, текстовые модели LLaMA (zero-shot и fine-tuned), мультимодальный компонент SALMONN-7B. Применение адаптивных функций потерь (Weighted Focal Loss, Vector Scaling) является необходимым условием для работы с выраженно несбалансированным датасетом MSP-PODCAST [1, с.3–4]. Несмотря на значительную архитектурную сложность, итоговый “state-of-the-art" результат составляет лишь 44.02% macro-F1 [1, с.4], что демонстрирует масштаб проблемы при переходе от лабораторных к реальным данным.

Работа [32] представляет направление, нацеленное на эффективное моделирование временных паттернов без необходимости чрезмерного усложнения архитектуры. Модель NCDE, использующая признаки Wav2vec2.0, достигает 73.37% WA на IEMOCAP при высокой стабильности и быстрой сходимости (72.88% WA уже на первой эпохе) [32, с.9–10]. Данный подход демонстрирует перспективность методов непрерывного моделирования, однако остаётся ограниченным качеством исходных данных, так как эффективность NCDE проявляется преимущественно на относительно чистых актерских наборах.

Одной из наиболее значимых проблем, обнаруживаемых в «натуралистичных» датасетах, является масштабный дисбаланс частот классов:

например, «нейтральные» эмоции встречаются в MSP-PODCAST в 26 раз чаще, чем «страх» [1, с.3]. Таким образом, методы коррекции дисбаланса являются неотъемлемой частью современных SER-систем, однако пока не существует универсальной стратегии для всех типов данных.

Результаты международного соревнования The Interspeech 2025 Challenge свидетельствуют о доминировании мультимодальных и ансамблевых решений: около 95% лучших систем используют комбинации аудио- и текстовых признаков, а также ансамблевые механизмы объединения предсказаний [29, с.4]. В частности, в системе "Abhinaya" простейшее голосование (majority voting) обеспечило относительный прирост производительности более чем на 24% по сравнению с лучшей одиночной моделью [1, с.4]. Тем не менее высокая вычислительная сложность и необходимость значительных аппаратных ресурсов делают такие системы менее пригодными для практических сценариев, особенно в мобильных и встроенных средах.

Можно сделать вывод, что анализ технологий SER демонстрирует чёткий разрыв между достижимыми результатами на контролируемых данных (~98% точности) и состоянием современных моделей на реальных, высокошумных данных (44.02% macro-F1) [25, с.12; 1, с.4]. Этот разрыв отражает фундаментальное противоречие между чистотой данных и практической применимостью.

Ключевые направления развития включают: переход к работе с «натуралистичными» данными, широкое внедрение фундаментальных моделей (Wav2vec2.0, WavLM, Whisper), увеличение роли мультимодальности как критического фактора качества, необходимость разработки менее ресурсоёмких, но устойчивых моделей. Дальнейшие исследования должны учитывать мультиязычность, доменную адаптацию и расширенную мультимодальность (аудио + видео), что позволит приблизить технологии SER к применению в реальных системах HCI [25, с.13].

1.3 Технологии анализа эмоций по тексту (NLP-based AER)

Современный ландшафт AER определяется доминированием трансформерных архитектур, прежде всего BERT и RoBERTa, чья способность моделировать тонкие контекстуальные зависимости выводит их далеко за пределы традиционных лексических и нейросетевых подходов [9, с.1]. В задачах многоязычного анализа особую роль играют модели XLM-RoBERTa, демонстрирующие благоприятный баланс между масштабируемостью и переносимостью [18, с.1; 17, с.1].

Несмотря на достигнутые результаты, современные AER-системы характеризуются рядом системных ограничений. Одной из ключевых проблем

остаётся интерпретация неявных эмоциональных сигналов, включая сарказм, иронию и сложные композиционные аффективные состояния. Такие феномены требуют не только лексической, но и дискурсивно-прагматической обработки, с чем базовые трансформеры справляются недостаточно эффективно [9, с.1]. Существенным остаётся и дисбаланс эмоциональных классов: миноритарные эмоции («страх», «удивление») оказываются выражены значительно слабее доминирующих категорий, что приводит к систематическому снижению полноты и ухудшению обобщающей способности [9, с.1; 28, с.1]. Третьим ограничением является недостаточная устойчивость к междоменных и межъязыковых сдвигам: модели, обученные на одном корпусе, нередко демонстрируют выраженное падение точности при переносе в новые лингвистические или культурные контексты, особенно в условиях низкоресурсных языков [17, с.1].

Анализ литературы позволяет выделить три стратегические линии развития AER. Первая связана с совершенствованием архитектур. Модель Emotion-Aware RoBERTa, предложенная в [9, с.1], дополняет трансформер специализированными механизмами Emotion-Specific Attention и TF-IDF–Gating, что обеспечивает повышение чувствительности к эмоционально насыщенным единицам текста и значимый прирост точности. Параллельно гибридные конструкции, например XLMCNN, объединяющая XLM-RoBERTa и сверточные слои, демонстрируют эффективность в задачах многоязычной классификации за счёт одновременного извлечения глобальных и локальных признаков [18, с.1].

Вторая линия развития связана с параметро-эффективной адаптацией. Включение адаптеров позволяет осуществлять тонкую настройку трансформеров без полного обновления весов, что существенно снижает вычислительные затраты и повышает устойчивость в низкоресурсных условиях. Результаты [17, с.1] показывают, что такие модели способны превосходить крупные LLM в задачах на языках с ограниченной аннотацией, включая тигринья и киньяруанда.

Третье направление заключается в оптимизации данных. Методика генерации пояснительных контекстов посредством LLM [21, с.1] позволяет значительно улучшить качество распознавания неявных эмоций за счёт включения семантически обогащённого входа. Традиционные методы аугментации, включая межъязыковой перевод, показали эффективность в увеличении разнообразия и снижении чувствительности к стилевым вариациям [5, с.1], хотя и не устраняют полностью проблему моделирования глубинных лингвистических особенностей.

Сопоставление рассмотренных подходов демонстрирует, что единое

универсальное решение отсутствует. Архитектурные усовершенствования обеспечивают максимальное качество в одноязычных задачах; параметро-эффективные методы оптимальны в условиях многоязычия и дефицита данных; подходы, основанные на обогащении данных, оказываются особенно результативны при работе с неоднозначными или контекстно-сложными текстами. Тем не менее, остаются нерешёнными вопросы повышения интерпретируемости, улучшения производительности для редких эмоциональных категорий и создания моделей, устойчивых к вариативности культурных и языковых сред [9, с.1; 17, с.1; 28, с.1].

1.4 Мультимодальное распознавание эмоций в человеко-компьютерном взаимодействии (HCI)

Мультимодальное распознавание эмоций (Multimodal Emotion Recognition, MER) играет ключевую роль в современном HCI, обеспечивая более естественную и адаптивную коммуникацию человека с цифровыми системами. В отличие от унимодальных методов, ориентированных на отдельные сигналы, мультимодальные подходы объединяют визуальные, акустические и текстовые данные, что приводит к более надежному и контекстно обоснованному пониманию эмоциональных состояний пользователей [33, с.1–2].

Анализ современных исследований демонстрирует постепенный переход от ранних архитектур с простым слиянием признаков к моделям глубокого обучения и, далее, к мультимодальным большим языковым моделям (MLLM), способным к высокоуровневому рассуждению об эмоциях [19, с.2].

Современный этап развития MER связан с появлением MLLM, которые интегрируют визуальные, текстовые и звуковые сигналы в едином пространстве представлений и способны не только классифицировать эмоции, но и объяснять свои решения на естественном языке [19, с.7]. Их применение в HCI принципиально меняет роль эмоций: от распознавания отдельных состояний к полноценному аффективному рассуждению. Существуют две основные стратегии адаптации MLLM к задачам MER.

Парадигма замороженных параметров (zero-shot и few-shot prompting) позволяет решать задачи без переобучения модели, что снижает вычислительные издержки и делает технологии доступными для реального HCI [19, с.9–10].

Парадигма настройки параметров (full tuning и PET/LoRA) обеспечивает точную адаптацию к целевому датасету, сохраняя эффективность за счёт минимальных обновляемых модулей [19, с.11–13].

Несмотря на значительный прогресс, MER в HCI сталкивается с

несколькими фундаментальными ограничениями:

1) недостаток крупномасштабных и репрезентативных мультимодальных наборов данных, что ограничивает способность моделей к генерализации [33, с.18];

2) асинхронность модальностей и необходимость тонкого выравнивания временных сигналов, особенно в диалоговых интерфейсах [19, с.24];

3) динамическая природа эмоций, не сводящихся к статичным меткам или кратким отрывкам данных [2, с.16];

4) субъективность аннотации эмоций и несоответствие между психологическими теориями и инженерными моделями [2, с.16];

5) необходимость интерпретируемых, лёгких моделей, адаптируемых для мобильных и носимых устройств — ключевых платформ HCI [33, с.17].

Анализ литературы показывает, что мультимодальное распознавание эмоций в HCI прошло путь от простых схем слияния признаков к сложным моделям глубокого обучения, а затем — к мультимодальным LLM, способным к когнитивно ориентированному рассуждению об эмоциях. Ключевым становится не просто повышение точности классификации, а интеграция контекста, моделирование динамики эмоциональных процессов и обеспечение интерпретируемости систем. Современные исследования указывают на необходимость объединения вычислительных методов с психологически обоснованными моделями и создания высококачественных мультимодальных ресурсов, что станет основой для следующего поколения эмоционально-интеллектуальных интерфейсов.

1.5 Анализ существующих решений и систем в человеко-компьютерном взаимодействии

Современные системы человеко-компьютерного взаимодействия (Human-Computer Interaction, HCI) переживают качественный технологический сдвиг, обусловленный быстрым развитием крупных языковых моделей (LLM) и мультимодальных моделей (MLLM). Чат-боты, голосовые ассистенты и интерфейсы, учитывающие эмоции, эволюционируют от систем, выполняющих фиксированные команды, к интеллектуальным агентам, способным интерпретировать контекст и эмоциональное состояние пользователя.

В данном разделе проводится аналитический анализ ключевых архитектурных подходов и нерешённых проблем, определяющих текущее состояние HCI.

Ключевым направлением развития HCI является создание эмоционально-чувствительных систем. Мультимодальное распознавание эмоций (Multimodal Emotion Recognition, MER) рассматривается в литературе как

фундаментальный механизм, обеспечивающий интеграцию речи, текста, мимики и физиологических сигналов [24, с.4; 33, с.1]. Мультимодальность обеспечивает устойчивость при отсутствии отдельных каналов и снижает неоднозначность, характерную для унимодальных решений [33, с.1].

Новый этап связан с использованием крупных предобученных моделей. В частности, CLIP позволил реализовать две разные стратегии слияния: визуально-якорную (VEGA) [15, с.1–2] и текстово-якорную (MER-CLIP) [24, с.2–4]. Первая ближе к перцептивному восприятию эмоций, вторая — более гибка в работе с лингвистическими обозначениями. Эти различия определяют направление проектирования будущих HCI-систем — либо перцептивно-универсальных, либо семантически-адаптивных.

Анализ работ указывает на несколько фундаментальных проблем, тормозящих внедрение эмоционально-интеллектуальных HCI:

1) Эмоциональные сигналы часто асинхронны, а современные MLLM недостаточно хорошо моделируют микродинамику и каузальные связи во времени [19, с.24–25; 2, с.16].

2) Одинаковые аудиовизуальные признаки могут соответствовать различным эмоциям в зависимости от контекста, что создаёт «семантический разрыв» [2, с.16].

3) Нехватка масштабных, сбалансированных датасетов снижает обобщающую способность моделей и приводит к сильной зависимости от конкретного корпуса [33, с.18].

4) Большие модели сложно применять в реальном времени, а отсутствие объяснимости ограничивает внедрение в критически важных сферах [33, с.17].

Актуальные направления исследований включают иерархические и причинно-следственные архитектуры [19, с.25–26], психологически обоснованные представления эмоций [2, с.17], самообучение [33, с.19], а также параметро-эффективные методы адаптации.

Анализ существующих HCI-систем демонстрирует глубокую трансформацию, связанную с переходом от текстовых моделей к мультимодальным архитектурам. MER становится ключевой технологией для построения эмоционально-чувствительных интерфейсов, а CLIP и MLLM — основой для преодоления ограничений традиционных методов. Несмотря на существенный прогресс, нерешёнными остаются вопросы межмодального выравнивания, недостатка данных и интерпретируемости, что определяет повестку будущих исследований в области HCI.

1.6. Проблемы и ограничения современных информационных технологий распознавания эмоций

Современные технологии FER, SER, AER и мультимодального MER демонстрируют значительный прогресс, однако сопоставительный анализ указывает на фундаментальные ограничения, возникающие независимо от используемой модальности и архитектурного класса моделей. Эти ограничения носят системный характер и отражают несоответствие между статистической природой методов глубокого обучения и многомерной, контекстуально обусловленной природой эмоциональной коммуникации.

1) Ограниченная обобщающая способность моделей.

Общий для всех модальностей эффект — резкое расхождение между результатами на контролируемых датасетах и качеством работы в реальных («in-the-wild») условиях. В FER модели, превосходящие 90% точности на CK+ или RAF-DB, теряют устойчивость при окклюзиях, изменении позы и освещения. В SER разрыв наиболее выражен: от ~98% на актёрских данных до ~44% macro-F1 на естественных записях. В AER трансформерные модели ухудшают качество при переносе на новые языки и домены.

Общей причиной является зависимость современных архитектур от стабильных статистических закономерностей, что ограничивает их способность обрабатывать вариативность поведения в реальной коммуникации. Усложнение архитектур одновременно повышает вычислительные затраты, но не снимает фундаментальной проблемы слабой способности к обобщению.

2) Недостаточность семантического и каузального моделирования.

Сравнение подходов показывает, что архитектурные инновации (attention-механизмы, SSL, MLLM) лишь частично компенсируют отсутствие у моделей способности к интерпретации причинно-следственных связей: FER модели фиксированы на визуальных паттернах и не учитывают ситуационный контекст; SER модели слабо различают эмоции, имеющие схожие акустические характеристики, но различающийся прагматический смысл; AER модели остаются уязвимыми к имплицитным эмоциям, иронии и полисемии. Даже мультимодальные MLLM, способные генерировать текстовые объяснения, в значительной степени выполняют пост-hoc интерпретацию, не отражающую реальных механизмов классификации.

3) Асинхронность и дисбаланс модальностей в MER.

Несмотря на распространённое мнение о преимущественности мультимодальности, сравнительный анализ показывает, что интеграция слабых или шумных сигналов приводит к деградации качества. Типичные проблемы включают: временную несогласованность аудио-, видео- и текстовых потоков; взаимное усиление шума; зависимость итоговой производительности от наиболее слабой модальности. Таким образом, мультимодальность улучшает качество только при условии надёжности каждого отдельного канала, что

существенно ограничивает применение MER в реальных интерфейсах.

4) Недостаточная интерпретируемость и практическая применимость.

Большинство современных моделей, включая CNN-гибриды, SSL и трансформеры, остаются «чёрными ящиками», что затрудняет их использование в критически важных областях. Сравнительный анализ показывает: landmark-ориентированные FER-модели обеспечивают наибольшую интерпретируемость, но уступают в универсальности; аудио- и текстовые модели обладают высокой точностью, но практически не поддаются анализу; MLLM предоставляют объяснения, не являющиеся частью процесса рассуждения модели.

Отсутствие прозрачности препятствует внедрению эмоционально-чувствительных систем в медицину, образование и социальные сервисы.

5) Культурная предвзятость и нерепрезентативность данных.

Все типы систем демонстрируют зависимость от распределения данных: FER модели хуже работают на этнически неоднородных выборках; SER модели — на многоязычных данных; AER — при переносе на незнакомые дискурсивные практики; MER — наследуют предвзятость каждой модальности одновременно.

Структурная проблема заключается в отсутствии универсальной модели эмоций: современные ИТ воспроизводят культурно специфические сигналы, а не интерпретируют эмоции как универсальное явление.

6) Ограниченная эффективность архитектурных инноваций

Сравнение различных подходов показывает, что прирост точности от усложнения архитектур постепенно уменьшается. Интересно, что: семантически направленные модели FER (ориентиры, внимание к ключевым зонам) зачастую превосходят более тяжёлые архитектуры Transformer; в SER фундаментальные SSL-модели обеспечивают скачок качества, но сильно зависят от «чистоты» данных; в AER локально адаптированные и параметро-эффективные модели могут превосходить крупные LLM.

Таким образом, дальнейшее увеличение размера моделей не обеспечивает пропорционального роста качества и ведёт к росту вычислительной стоимости.

7) Технологические ограничения и отсутствие лёгких систем для реального времени.

Большинство современных моделей характеризуются высокими требованиями к аппаратным ресурсам: FER-модели достигают 5–10 GFLOPs; SSL-аудиомодели требуют значительных ресурсов для инференса; мультимодальные модели (CLIP, MLLM) оперируют сотнями миллионов или миллиардами параметров. Это ограничивает их применение.

ГЛАВА 2
МЕТОДИКА ИССЛЕДОВАНИЯ

В рамках данного исследования применяется комплексный методологический подход, сочетающий систематический литературный анализ, критическую оценку современных архитектур нейросетевых моделей и методы сравнительного анализа технологий, используемых в задачах распознавания эмоций по визуальным, аудио, текстовым и мультимодальным данным.

Основная цель методики — обеспечить структурированное и объективное представление текущего состояния области, выявить ключевые технические вызовы и определить перспективные направления развития эмоционально-интеллектуальных систем в контексте человеко-компьютерного взаимодействия (HCI).

Источники анализировались по нескольким уровням критериев, отражающих современные требования к системам распознавания эмоций. Первый критерий это точность и стабильность моделей, которые влияют на качество классификации эмоций в различных условиях (освещение, шум, окклюзия, вариативность речи, многокультурность данных). Далее способность моделей работать на разных датасетах и обрабатывать “in-the-wild" данные. Кроме того, большое внимание было уделено пригодности архитектур для edge-вычислений, мобильных устройств и реального времени. Также важным критерием для доверия пользователя и корректного внедрения в HCI-системы является интерпретируемость и прозрачность работы моделей.

С целью обеспечения последовательности анализа была разработана структурированная схема классификации источников. Все рассмотренные публикации распределены по основным технологическим направлениям: FER (visual emotion recognition), SER (speech emotion recognition), AER (text-based emotion analysis), мультимодальные методы, а также работы, посвящённые анализу существующих решений. Такое распределение позволяет проводить сравнение методов как внутри отдельных доменов, так и между ними, выявляя общие тенденции и специфические проблемы каждого направления.

В исследовании использовались также практические инструменты анализа: тестирование моделей на открытых датасетах FER и SER (например, RAF-DB, AffectNet, EmoDB), экспериментальные проверки устойчивости к шуму и изменениям входных данных, а также сравнение результатов разных архитектур в средах PyTorch и TensorFlow. Это позволяло оценивать не только теоретические выводы авторов, но и их применимость к реальным задачам.

Таким образом, предложенная методика позволила не только обобщить текущее состояние технологий распознавания эмоций, но и определить перспективные методы, подходящие для современных HCI-систем.

ГЛАВА 3
РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Проведённое исследование позволило выявить совокупность закономерностей, определяющих текущее состояние технологий автоматического распознавания эмоций и их применимость в современных человеко-компьютерных интерфейсах. Анализ существующих решений показал, что ключевым фактором, влияющим на эффективность всех модальностей — визуальной, аудио, текстовой и мультимодальной, — остаётся разрыв между лабораторными условиями и реальными сценариями использования. Именно при переходе к данным “in-the-wild” происходит наиболее заметная деградация качества. Для систем распознавания по голосу это выражается в снижении метрик с почти идеальных значений на актёрских записях до уровней, близких к случайному угадыванию, при работе со спонтанной речью в условиях шума. Аналогичное явление наблюдается в анализе лицевых выражений: даже модели, демонстрирующие высокую точность в контролируемых условиях, существенно теряют робастность при наличии окклюзий, изменении ракурса или освещённости. Причина заключается в высокой зависимости моделей от статистически стабильных шаблонов, характерных для искусственных датасетов, и их ограниченной способности к обобщению.

Выявлено также, что дальнейшее увеличение архитектурной сложности уже не приводит к пропорциональному росту качества. Более того, сравнительный анализ показал преимущество моделей, использующих семантически обоснованные представления данных, над сетями, полагающимися исключительно на высокую емкость. Так, методы, основанные на лицевых ориентирах, обеспечивают более устойчивый результат, чем глубокие гибридные сети, особенно в условиях ограниченных вычислительных ресурсов. Похожая картина наблюдается и в текстовой модальности: параметро-эффективные подходы адаптации, такие как адаптеры или LoRA, демонстрируют сопоставимое или более высокое качество по сравнению с полным дообучением больших языковых моделей, сохраняя при этом значительно меньшие требования к ресурсам.

Отдельное внимание в ходе исследования уделено роли данных. Показано, что качество разметки и репрезентативность датасетов оказывают более значимое влияние на итоговую точность, чем выбор конкретной архитектуры. Распространённые проблемы, включая дисбаланс классов и культурную предвзятость, формируют систематические ошибки, которые не могут быть устранены исключительно за счёт увеличения объёма моделей. Особенно критичным остаётся дефицит масштабных, разнообразных и

согласованных мультимодальных наборов данных, необходимых для полноценного развития мультимодальных систем распознавания эмоций.

Несмотря на очевидные преимущества мультимодальных подходов, исследование показало, что синхронизация разнородных каналов (визуального, аудио и текстового) является нетривиальной задачей, и в реальных условиях мультимодальные модели нередко наследуют слабые стороны своих отдельных компонент. В результате комбинация каналов приносит выгоду лишь тогда, когда каждый из них обладает достаточным качеством; в противном случае более слабая модальность способна ухудшить итоговый результат. Это ограничивает применение мультимодальных систем в критически значимых задачах и требует разработки адаптивных схем фьюжна, учитывающих качество и доступность отдельных сигналов в реальном времени.

Практическая значимость полученных результатов заключается в формировании рекомендаций по созданию устойчивых и интерпретируемых систем распознавания эмоций для HCI. Исследование показало, что успешные решения строятся не на усложнении архитектур, а на комбинировании доменно информированных признаков, ориентировании на реальные условия работы и широком применении параметро-эффективных методов. Для визуальных систем наиболее перспективным подходом является использование семантически богатых представлений, основанных на геометрии лица. Для аудиальных систем — адаптация моделей к шумовым условиям и использование самообучения на больших неразмеченных корпусах речи. В текстовой модальности — применение лёгких методов адаптации к конкретным лингвистическим доменам. Что касается мультимодальных систем, их эффективность определяется модульным устройством, позволяющим отдельно контролировать вклад каждой модальности, а также интеграцией современных мультимодальных языковых моделей, обеспечивающих интерпретируемость и способность учитывать сложный контекст взаимодействия пользователя с системой.

В результате сформирована комплексная картина отрасли: от выявленных ограничений и системных уязвимостей до перспективных направлений развития. К ним относятся разработка моделей, оптимизированных для мобильных устройств и реального времени; использование психологически валидированных теоретических структур эмоций в качестве основы разметки и обучения; расширение практики самообучения; а также внедрение мультимодальных языковых моделей, способных рассуждать о намерениях, состояниях и поведенческих паттернах пользователя. Совокупность этих подходов определяет стратегию перехода к новому поколению эмоционально-интеллектуальных интерфейсов.

ЗАКЛЮЧЕНИЕ

В данной работе было проведено комплексное исследование современных информационных технологий распознавания эмоций в контексте человеко-компьютерного взаимодействия (HCI).

В первой главе проведён критический обзор ключевых архитектурных подходов в области распознавания эмоций. Выявлены фундаментальные технологические барьеры, такие как разрыв в производительности моделей между лабораторными и реальными условиями («in-the-wild»), чувствительность к шуму, окклюзиям и изменению контекста, а также проблемы дисбаланса классов и культурной предвзятости.

Во второй главе описана методика исследования, включающая систематический литературный анализ и сравнительную оценку нейросетевых архитектур по критериям точности, обобщающей способности и применимости в реальных условиях HCI. Было показано, что важнейшими факторами качества систем распознавания эмоций являются не только архитектурные инновации, но и семантически-обоснованные признаки, а также адаптивные и параметро-эффективные подходы.

В третьей главе сформулирован центральный вывод работы: неконтролируемое усложнение архитектур не обеспечивает пропорционального роста качества распознавания. На практике более устойчивыми и интерпретируемыми оказываются модели с семантически-направленным извлечением признаков и подходы, оптимизированные для конкретной доменной или мультимодальной задачи. Анализ показал, что мультимодальные системы обеспечивают значительный потенциал, однако их эффективность ограничена синхронизацией и качеством отдельных каналов.

Таким образом, результаты работы позволяют определить стратегические направления дальнейшего развития эмоционально-интеллектуальных интерфейсов: создание компактных и адаптивных моделей для мобильных и реального времени, внедрение психологически обоснованных структур эмоций для обучения моделей, широкое использование параметро-эффективных методов и постепенная интеграция мультимодальных языковых моделей, способных к интерпретируемому аффективному рассуждению.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. ABHINAYA -- A System for Speech Emotion Recognition In Naturalistic Conditions Challenge [Электронный ресурс] / S. Dutta [et al.] // arXiv.org. – 2025. – Режим доступа: https://arxiv.org/abs/2505.18217. – Дата доступа: 01.11.2025.

2. Advances in Video Emotion Recognition: Challenges and Trends / Y. Yi [et al.] // Sensors. – 2025. – Vol. 25. – P. 3615.

3. Almubarak, M. An Ensemble Learning Approach for Facial Emotion Recognition Based on Deep Learning Techniques / M. Almubarak, F. A. Alsulaiman // Electronics. – 2025. – Vol. 14. – P. 3415.

4. Ankomah, E. Emotion-Aware AI Chatbots for Mental Health Support in Low-Resource Public Health Systems: A Case Study from Ghana / E. Ankomah, R. E. Turkson // World Journal of Public Health. – 2025. – Vol. 10, № 3. – P. 17.

5. Aruna Gladys, A. Survey on multimodal approaches to emotion recognition / A. Aruna Gladys, V. Vetriselvi // Neurocomputing. – 2023. – Vol. 556. – P. 126693.

6. Baltrušaitis, T. Multimodal Machine Learning: A Survey and Taxonomy [Электронный ресурс] / T. Baltrušaitis, C. Ahuja, L.-P. Morency // arXiv.org. – 2017. – Режим доступа: https://arxiv.org/abs/1705.09406. – Дата доступа: 20.11.2025.

7. Egan, L. A. ReNeuWell mental well-being app: protocol for a randomised controlled trial / L. A. Egan, J. M. Gatt // BMJ Open. – 2025. – Vol. 15, № 4. – P. e094557.

8. Emotion-Aware Chatbot Architecture: Enhancing Human-Robot Interaction through Sentiment Detection and Lip Sync / M. Abdelaziz [et al.] // Proceedings of the 5th Biennial African Human Computer Interaction Conference (AfriCHI '25). – New York : Association for Computing Machinery, 2025. – P. 474–477.

9. Emotion-Aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition / F. Alqarni [et al.] // Scientific Reports. – 2025. – Vol. 15. – P. 17617.

10. Enhanced Facial Emotion Recognition and Age Estimation Using Modified Residual Network and Support Vector Machine / N. A. El-Hag [et al.] // International Journal of Computational Intelligence Systems. – 2025. – Vol. 18. – P. 231.

11. Evaluating a brief smartphone-based stress management intervention with heart rate biofeedback from built-in sensors in a three arm randomized controlled trial / L. M. Fuhrmann [et al.] // Scientific Reports. – 2025. – Vol. 15. – P. 20257.

12. Gartner: 40% of GenAI Solutions Will Be Multimodal by 2027 [Электронный ресурс]. – 2024. – Режим доступа: https://www.apmdigest.com/gartner-40-genai-solutions-will-be-multimodal-2027. – Дата доступа: 16.11.2025.

13. Goswami, S. The Need for Emotional Intelligence in Human-Computer Interactions / S. Goswami, S. Dave, K. Patel // Multidisciplinary Approaches in Affective Computing. – Hershey, PA : IGI Global, 2024. – P. 104–123.

14. Gutierrez, R. Development of adaptive and emotionally intelligent educational assistants based on conversational AI / R. Gutierrez, W. Villegas-Ch, J. Govea // Frontiers in Computer Science. – 2025. – Vol. 7. – P. 1628104.

15. Hu, G. Grounding Emotion Recognition with Visual Prototypes: VEGA - Revisiting CLIP in MERC / G. Hu, D. Kollias, X. Yang // Proceedings of the 33rd ACM International Conference on Multimedia (MM '25). – New York : Association for Computing Machinery, 2025. – P. 5667–5676.

16. Human–computer interaction for cognitive, emotional and learning well-being / K. Upreti [et al.] // Intelligent Systems for Neurocognition and Human-Robot-Computer Interaction / ed. by S. Mahajan, D. S. Kapoor, K. J. Singh. – London : Academic Press, 2026. – P. 43–65.

17. Laureano, F. UoB-NLP at SemEval-2025 Task 11: Leveraging Adapters for Multilingual and Cross-Lingual Emotion Detection [Электронный ресурс] / F. Laureano, Y. Wang, Y. Feng // arXiv.org. – 2025. – Режим доступа: https://arxiv.org/abs/2504.08543. – Дата доступа: 01.11.2025.

18. Li, J. EMO-NLP at SemEval-2025 Task 11: Multi-label Emotion Detection in Multiple Languages Based on XLM-CNN / J. Li, Y. Xian, X. Yang // Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). – Vienna : Association for Computational Linguistics, 2025. – P. 801–806.

19. Multimodal Large Language Models Meet Multimodal Emotion Recognition and Reasoning: A Survey / Y. Shou [et al.] // Journal of the ACM. – 2025. – Vol. 1, № 1.

20. Pantic, M. Toward an affect-sensitive multimodal human-computer interaction / M. Pantic, L. J. M. Rothkrantz // Proceedings of the IEEE. – 2003. – Vol. 91, № 9. – P. 1370–1390.

21. Ranjbar, N. Lotus at SemEval-2025 Task 11: RoBERTa with Llama-3 Generated Explanations for Multi-Label Emotion Classification / N. Ranjbar, H. Baghbani // Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). – Vienna : Association for Computational Linguistics, 2025. – P. 431–439.

22. So, J. Facial Landmark-Driven Keypoint Feature Extraction for Robust Facial Expression Recognition / J. So, Y. Han // Sensors. – 2025. – Vol. 25. – P. 3762.

23. Song, D. A facial expression recognition network using hybrid feature extraction / D. Song, C. Liu // PLoS ONE. – 2025. – Vol. 20, № 1. – P. e0312359.

24. Song, Y. Leveraging CLIP Encoder for Multimodal Emotion Recognition /

Y. Song, S. Cho // Proceedings of the Winter Conference on Applications of Computer Vision (WACV). – Waikoloa : IEEE, 2025. – P. 6115–6124.

25. Speech Emotion Recognition: Comparative Analysis of CNN-LSTM and Attention-Enhanced CNN-LSTM Models / J. Bhanbhro [et al.] // Signals. – 2025. – Vol. 6. – P. 22.

26. Systematic Review of Emotion Detection with Computer Vision and Deep Learning / R. Pereira [et al.] // Sensors. – 2024. – Vol. 24. – P. 3484.

27. Taiwo, O. Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support / O. Taiwo, B. Al-Bander // Cognitive Computation and Systems. – 2025. – P. e12116.

28. Takenaka, Y. Performance Evaluation of Emotion Classification in Japanese Using RoBERTa and DeBERTa [Электронный ресурс] / Y. Takenaka // arXiv.org. – 2025. – Режим доступа: https://arxiv.org/abs/2505.00013. – Дата доступа: 01.11.2025.

29. The Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions / A. R. Naini [et al.] // Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2025). – Rotterdam, 2025. – P. 4668–4672.

30. Udahemuka, G. Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review / G. Udahemuka, K. Djouani, A. M. Kurien // Applied Sciences. – 2024. – Vol. 14. – P. 8071.

31. VerbaNexAI Lab at SemEval-2024 Task 3: Deciphering emotional causality in conversations using multimodal analysis approach / V. Pacheco [et al.] // Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). – Mexico City : Association for Computational Linguistics, 2024. – P. 1629–1635.

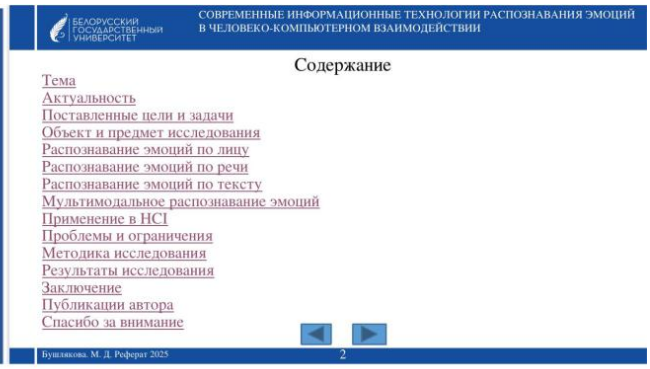
32. Wang, N. Speech emotion recognition using fine-tuned Wav2vec2.0 and neural controlled differential equations classifier / N. Wang, D. Yang // PLoS ONE. – 2025. – Vol. 20, № 2. – P. e0318297.

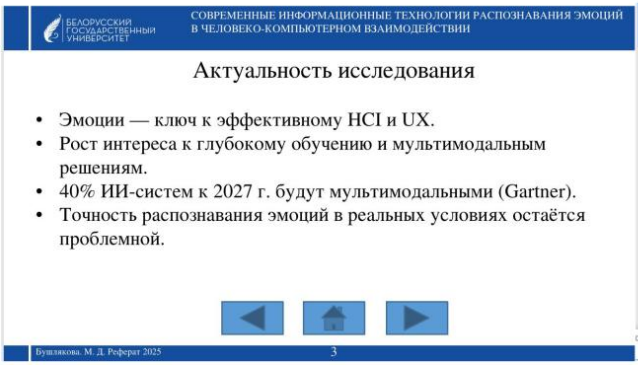
33. Wu, Y. A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions / Y. Wu, Q. Mi, T. Gao // Biomimetics. – 2025. – Vol. 10. – P. 418.

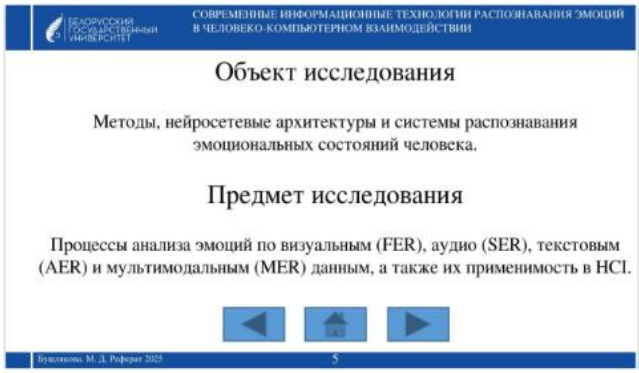
34. Zaheer, S. Designing Emotion-Aware UX: Leveraging Sentiment Analysis to Adapt Digital Experiences [Электронный ресурс] / S. Zaheer. – 2023. – Vol. 4, № 1. – Режим доступа: https://doi.org/10.5281/zenodo.15259144. – Дата доступа: 01.11.2025.

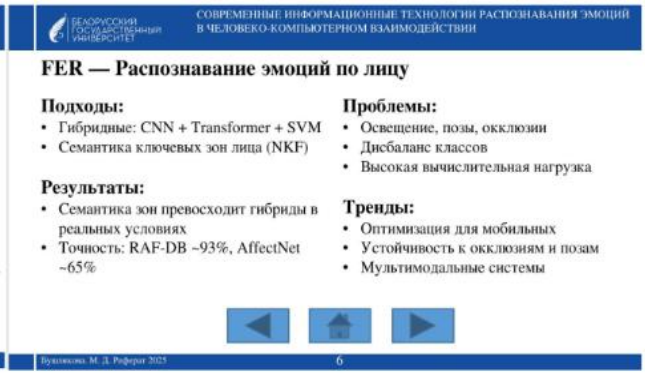
ПРИЛОЖЕНИЕ А ПРЕЗЕНТАЦИЯ РЕФЕРАТА

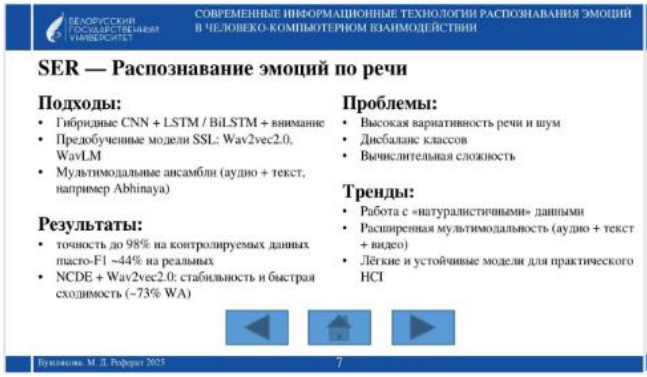
https://margoby.github.io/files/referat_presentation.pdf - ссылка на электронный вариант презентации

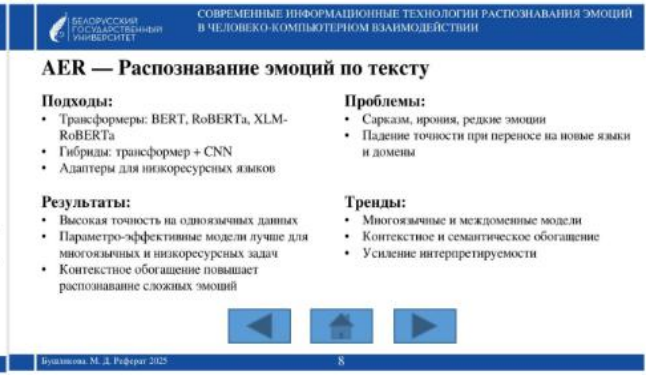












MER — Мультимодальное распознавание эмоций

Подходы:
- Объединение визуальных, аудио и текстовых данных
- Мультимодальные LLM: классификация + объяснение эмоций
- Адаптация: zero-/few-shot prompting или fine-tuning (PET/LoRA)

Проблемы:
- Недостаток крупномасштабных мультимодальных датасетов
- Асинхронность сигналов и динамика эмоций
- Интерпретируемость и адаптация к мобильным/носимым устройствам

Результаты:
- Более точное и контекстное понимание эмоций, чем у унимодальных моделей
- Поддержка когнитивного аффективного рассуждения в HCI

Тренды:
- Психологически обоснованные модели
- Контекстуализация и моделирование динамики эмоций
- Создание высококачественных мультимодальных ресурсов

HCI — Мультимодальное распознавание эмоций в человеко-компьютерном взаимодействии

Подходы:
- Эмоционально-чувствительные интерфейсы: MER + MLLM
- Слияние визуальных, аудио и текстовых данных (CLIP: VEGA vs MER-CLIP)
- Параметро-эффективная настройка и самообучение

Проблемы:
- Асинхронность сигналов и микродинамика эмоций
- «Семантический разрыв» контекстных эмоций
- Нехватка масштабных сбалансированных датасетов
- Ограниченная интерпретируемость и высокая вычислительная сложность

Результаты:
- Устойчивость к отсутствию отдельных каналов
- Более точная интерпретация эмоций и контекста
- Поддержка мультимодальных и семантически адаптивных систем

Тренды:
- Иерархические и причинно-следственные архитектуры
- Психологически обоснованные представления эмоций
- Интеграция MER с крупными мультимодальными моделями

Проблемы и ограничения современных ИТ распознавания эмоций

Общие ограничения:
- Высокие показатели на контролируемых данных vs низкая точность «in-the-wild»
- Недостаточное семантическое и каузальное моделирование
- Асинхронность и дисбаланс модальностей в MER
- Ограниченная интерпретируемость и практическая применимость
- Культурная предвзятость и нерепрезентативность данных

Архитектурные и технологические ограничения:
- Дальнейшее усложнение моделей не гарантирует рост качества
- Высокие вычислительные затраты: FER ~5–10 GFLOPs, мультимодальные MLLM сотни млн–млрд параметров
- Отсутствие лёгких и мобильных решений для реального времени

Вывод:

Прогресс достигнут, но фундаментальные проблемы «черного ящика», вариативности данных и мультимодальной интеграции остаются нерешёнными

Методика исследования

Подход:
- Систематический литературный обзор и критический анализ нейросетевых архитектур
- Сравнительный анализ технологий FER, SER, AER и мультимодальных методов

Ключевые критерии оценки:
- Точность и стабильность моделей («in-the-wild», шум, окклюзии, вариативность речи, культурная разнородность)
- Универсальность и обобщаемость на разных датасетах
- Поддержка мобильных устройств и edge-вычислений
- Интерпретируемость и прозрачность работы моделей

Цель:
- Объективное представление текущего состояния технологий
- Выявление ключевых вызовов и перспектив для эмоционально-интеллектуальных HCI-систем

Результаты исследования

Ключевые выводы:
- Разрыв между лабораторными и реальными данными (существенное падение точности всех модальностей)
- Усложнение архитектур не гарантирует рост качества
- Семантически обоснованные признаки и параметро-эффективные методы часто превосходят тяжёлые модели
- Качество и репрезентативность данных важнее архитектуры (дисбаланс классов и культурная предвзятость вызывает систематические ошибки)

Особенности мультимодальности:
- Выгода при высоком качестве всех каналов
- Слабая модальность может ухудшать итоговый результат
- Требуются адаптивные схемы фьюжна и контроль вкладов каждого канала

Результаты исследования

Практическая значимость:
- Визуальные системы: семантически богатые лицевые признаки
- Аудио: адаптация к шуму и самообучение на больших корпусах
- Текст: лёгкие методы адаптации к домену
- Мультимодальные системы: модульная структура + интеграция MLLM для интерпретируемости

Перспективы развития:
- Оптимизация моделей для мобильных устройств и реального времени
- Использование психологически валидированных теорий эмоций
- Расширение самообучения
- Мультимодальные языковые модели для контекстного и когнитивного распознавания эмоций

Заключение

- Проведён комплексный анализ технологий распознавания эмоций
- Выявлены ключевые проблемы: падение качества «in-the-wild», чувствительность к шуму, окклюзиям, дисбаланс классов и культурная предвзятость
- Сформулированы основные направления развития информационных технологий распознования эмоций для HCI: компактные модели для мобильных и реального времени, использование психологически обоснованных структур эмоций, параметро-эффективная адаптация, интеграция мультимодальных языковых моделей с интерпретируемым аффективным рассуждением

Публикации автора

Бушлякова М. Д., Использование сред моделирования в процессе изучения робототехники и алгоритмов машинного обучения / М.Д. Бушлякова // Инновационные подходы к обучению физике, математике, информатике : материалы Междунар. студ. науч.-практ. конф., г. Минск, 27 марта 2025 г. / Белорус. гос. пед. ун-т им. М. Танка; редкол. В. В. Радыгина, А. А. Франкевич (отв. ред.) [и др.]. – Минск : БГПУ, 2025.

СПАСИБО ЗА ВНИМАНИЕ

ПРИЛОЖЕНИЕ Б ПЕРСОНАЛЬНЫЙ САЙТ

Адрес сайта: https://margoby.github.io/

Бушлякова Маргарита Дмитриевна

Магистрант 1 курса
Белорусский государственный университет
Факультет **ФПМИ**, кафедра **ИСУ**

Образование

- 2025–2027 — магистратура, БГУ, ФПМИ, кафедра ИСУ, специальность «Прикладная математика и информатика»
- 2021–2025 — бакалавриат, БГУ, ФПМИ, кафедра КТС, специальность «Информатика»

Научные интересы

- Машинное обучение
- Мультимодальный анализ данных
- Обучение с подкреплением

Публикации

- Бушлякова М. Д., Использование сред моделирования в процессе изучения робототехники и алгоритмов машинного обучения // Инновационные подходы к обучению физике, математике, информатике : материалы Междунар. студ. науч.-практ. конф., г. Минск, 27 марта 2025 г. / Белорус. гос. пед. ун-т им. М. Танка; редкол. В. В. Радыгина, А. А. Францкевич (отв. ред.) [и др.]. – Минск : БГПУ, 2025.

Работы и рефераты

Реферат по КДЗ «Основы информационных технологий»

Скачать реферат: referat.pdf

Презентация: referat_presentation.pdf

Дипломная работа

Скачать работу: diploma.pdf

Контакты

Email: margo.bushliakova@gmail.com
GitHub: https://github.com/margoby