УДК 004.93

Крикунов Данила Анатольевич, магистрант, Тюменский Индустриальный Университет, г. Тюмень
Машаров Михаил Максимович, магистрант, Тюменский Индустриальный Университет, г. Тюмень

СИНТЕЗ МУЛЬТИМОДАЛЬНЫХ ПРИЗНАКОВ В ЗАДАЧАХ КЛАССИФИКАЦИИ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ СУБЪЕКТОВ КОММУНИКАЦИИ

Аннотация

В статье рассматривается задача повышения точности классификации эмоционального состояния субъектов коммуникации на основе синтеза мультимодальных признаков, извлекаемых из текстовых, акустических и визуальных данных. Актуальность исследования обусловлена ограничениями одномодальных систем анализа эмоций, демонстрирующих снижение качества распознавания в условиях шумов, неполноты информации, асинхронности каналов и вариативности речевого поведения. В работе предлагается подход к адаптивной мультимодальной фузии, учитывающий качество отдельных модальностей, временную синхронизацию признаков и степень достоверности результатов распознавания.

Исследование выполняется на примере архитектуры программной системы ComIntellect, предназначенной для автоматизированного анализа человеческих коммуникаций. Рассматриваются этапы подготовки медиаданных, Voice Activity Detection, диаризации спикеров, автоматического распознавания речи, сопоставления слов с участниками диалога и мультимодального обогащения реплик. Предложена математическая модель динамического взвешивания признаков, позволяющая формировать единое

пространство представления эмоционального состояния на основе текстовых, акустических и визуальных сигналов.

Отдельное внимание уделяется применению больших языковых моделей для задач семантической суммаризации, интерпретации эмоциональной динамики и генерации evidence-based рекомендаций. Показано, что использование адаптивной мультимодальной фузии позволяет повысить устойчивость классификации эмоциональных состояний и обеспечить более точный анализ коммуникационного поведения в сравнении с одномодальными подходами.

Annotation

The article considers the problem of improving the accuracy of emotional state classification of communication subjects based on the synthesis of multimodal features extracted from textual, acoustic and visual data. The relevance of the study is обусловлена the limitations of unimodal emotion analysis systems, which demonstrate reduced recognition quality under conditions of noise, incomplete information, asynchronous data streams and variability of speech behavior. The paper proposes an approach to adaptive multimodal fusion that takes into account the quality of individual modalities, temporal synchronization of features and the confidence level of recognition results.

The study is carried out using the architecture of the ComIntellect software system intended for automated analysis of human communications. The paper examines the stages of media preparation, Voice Activity Detection, speaker diarization, automatic speech recognition, word-to-speaker alignment and multimodal utterance enrichment. A mathematical model of dynamic feature weighting is proposed, allowing the formation of a unified representation space of emotional states based on textual, acoustic and visual signals.

Special attention is paid to the use of large language models for semantic summarization, interpretation of emotional dynamics and generation of evidence-based recommendations. It is shown that the use of adaptive multimodal fusion increases the robustness of emotional state classification and provides more accurate communication behavior analysis compared to unimodal approaches.

Ключевые слова: мультимодальная фузия, классификация эмоций, анализ коммуникаций, диаризация спикеров, темпоральная синхронизация признаков, асинхронный инференс, LLM, семантическая суммаризация
Keywords: multimodal fusion, emotion classification, communication analysis, speaker diarization, temporal feature synchronization, asynchronous inference, large language models, semantic summarization

Современные цифровые системы анализа коммуникаций ориентируются на автоматизированную обработку эмоционального состояния субъектов взаимодействия, поскольку качество коммуникации напрямую влияет на эффективность переговоров, клиентского сервиса, дистанционного обучения и управления конфликтными ситуациями. В исследованиях, посвященных мультимодальному восприятию эмоций, подчеркивается, что эмоциональное состояние человека формируется одновременно несколькими каналами коммуникации: речевыми, визуальными и семантическими [1]. В условиях роста объема аудио- и видеокоммуникаций ручной анализ эмоционального поведения становится трудоемким, субъективным и плохо масштабируемым процессом. В связи с этим особую актуальность приобретает разработка интеллектуальных систем, способных автоматически интерпретировать эмоциональную динамику участников диалога на основе комплексной обработки мультимодальных данных.

Традиционные методы классификации эмоционального состояния преимущественно основываются на анализе одной модальности: текста, аудиосигнала либо визуальных признаков. Однако эмоциональное поведение человека формируется совокупностью взаимосвязанных факторов, включающих лингвистическое содержание высказывания, просодические характеристики речи, темп коммуникации, наличие пауз, интенсивность интонации, мимику и микровыражения лица. Использование только одного канала анализа приводит к существенному снижению точности распознавания эмоций в условиях шумов, неполноты данных и асинхронности поступления

сигналов. Акустические методы классификации эмоций демонстрируют высокую зависимость от качества речевого сигнала и чувствительны к наложению голосов и фоновым шумам [2].

Текстовые модели анализа эмоций демонстрируют высокую зависимость от качества автоматического распознавания речи и не способны корректно интерпретировать эмоциональную окраску высказываний при наличии сарказма, контекстной неоднозначности и фрагментарности диалога. Визуальные методы классификации эмоций также имеют ряд ограничений, связанных с качеством видеопотока, освещением, углом обзора и частичным перекрытием лица. Исследования в области компьютерного зрения показывают, что точность распознавания мимических реакций существенно снижается в условиях нестабильного видеосигнала [4]. Вследствие этого применение одномодальных систем в реальных коммуникационных сценариях не обеспечивает требуемой устойчивости классификации эмоциональных состояний.

Наиболее перспективным направлением развития интеллектуальных систем анализа коммуникаций является использование мультимодальной фузии признаков, позволяющей объединять текстовые, акустические и визуальные сигналы в единое пространство представления эмоционального состояния. Под мультимодальной фузией в данном случае понимается процесс синтеза разнородных признаков, извлекаемых из различных модальностей, с учетом временной синхронизации, качества каналов и контекста коммуникации. Современные исследования показывают, что комбинирование нескольких модальностей позволяет повысить устойчивость классификации эмоций по сравнению с одномодальными подходами [7]. Ключевой проблемой при построении подобных систем является необходимость обработки асинхронных потоков данных, содержащих различные интервалы времени и неоднородную структуру признакового пространства.

Дополнительную сложность представляет задача анализа многопользовательских коммуникаций, в которых присутствуют

перебивания, наложения речи, смена эмоционального состояния и неравномерная длительность реплик. В подобных условиях требуется не только определение эмоционального состояния, но и построение устойчивой архитектуры обработки данных, включающей Voice Activity Detection, диаризацию спикеров, автоматическое распознавание речи и механизм мультимодального обогащения реплик. В работах, посвященных emotion diarization, подчеркивается важность корректного разделения спикеров для последующего анализа эмоциональной динамики разговора [8].

В рамках исследования рассматривается архитектурный подход к построению системы мультимодального анализа коммуникаций, основанный на использовании модульного pipeline обработки аудио-, видео- и текстовых данных. На первом этапе выполняется подготовка медиаданных и нормализация входного сигнала. При обработке видео осуществляется извлечение аудиодорожки и унификация форматов представления данных. Далее применяется Voice Activity Detection, предназначенный для выделения участков речевой активности и исключения длинных сегментов тишины и фонового шума. Использование VAD позволяет снизить количество ложных распознаваний и уменьшить вычислительную нагрузку на последующие этапы анализа.

После выделения речевых сегментов выполняется задача диаризации спикеров. Для определения принадлежности реплик конкретным участникам диалога используется каскадная модель диаризации, включающая Voice Activity Detection, speaker embeddings и механизмы сегментации речевого потока. Диаризация играет критически важную роль в задачах анализа эмоционального поведения, поскольку позволяет связать эмоциональные признаки с конкретным субъектом коммуникации. В исследованиях по анализу эмоциональной диаризации речи отмечается, что корректное разделение реплик напрямую влияет на точность дальнейшей классификации эмоциональных состояний [5].

Следующим этапом pipeline является автоматическое распознавание речи. Для обработки речевого сигнала используются модели ASR, поддерживающие получение временных меток отдельных слов. Подобный подход обеспечивает возможность последующего сопоставления текстовых сегментов с результатами диаризации и временными интервалами эмоциональных сигналов. После получения расшифровки выполняется процедура сопоставления слов с конкретными участниками диалога, что позволяет формировать структурированные реплики с временными метками и идентификаторами спикеров.

После завершения сопоставления слов с участниками диалога формируются реплики, содержащие текст, временные интервалы и список слов с временными метками. Данная структура используется как базовый элемент дальнейшего мультимодального анализа. Для каждой реплики выполняется извлечение текстовых, акустических и визуальных признаков. Текстовые признаки включают семантические embeddings и результаты анализа эмоциональной окраски высказывания. Акустические признаки содержат спектральные характеристики сигнала, темп речи, интенсивность и длительность пауз. Визуальные признаки формируются на основе анализа мимики и facial embeddings. Подобный подход соответствует современным направлениям развития мультимодального sentiment analysis [7].

Ключевым этапом системы является мультимодальное обогащение реплик, в рамках которого все извлеченные признаки объединяются в единую структуру представления коммуникационного поведения. Помимо базовых эмоциональных характеристик в структуру анализа включаются дополнительные поведенческие сигналы: скорость речи, перебивания, overlap rate, длительность пауз и уровень эмоциональной интенсивности. Подобная структура позволяет перейти от анализа отдельных модальностей к комплексной интерпретации эмоционального состояния субъектов коммуникации.

Для повышения устойчивости классификации используется механизм адаптивной мультимодальной фузии признаков. В отличие от статических схем объединения, предполагающих одинаковую значимость всех каналов, адаптивный подход учитывает качество каждой модальности в текущий момент времени. При ухудшении качества аудиосигнала система автоматически снижает влияние акустических признаков и усиливает роль текстовых и визуальных данных. Аналогичным образом при отсутствии видеопотока или нестабильности распознавания лица увеличивается значимость других каналов. Подобный механизм позволяет повысить устойчивость анализа в условиях шумов, асинхронности данных и частичной потери информации [3].

Важной особенностью рассматриваемой архитектуры является темпоральная синхронизация признаков. Поскольку текстовые, аудио- и видеосигналы имеют различную частоту обновления и длительность интервалов, требуется механизм приведения признаков к единой временной шкале. Для решения данной задачи используются sliding window, timestamp alignment и interpolation-based synchronization. Темпоральная синхронизация позволяет корректно сопоставлять изменения эмоционального состояния с конкретными фрагментами коммуникации и предотвращает потерю контекстной информации.

Для классификации эмоционального состояния используется ансамбль моделей машинного обучения, включая градиентный бустинг и transformer-based classifiers. Применение градиентного бустинга представляет особый интерес, поскольку данный подход демонстрирует устойчивость при работе с разнородными признаками и обеспечивает интерпретируемость результатов. Итоговый классификатор формирует распределение вероятностей по классам эмоциональных состояний, включая нейтральное состояние, раздражение, тревогу, вовлеченность и эмоциональную напряженность.

Дополнительный уровень анализа обеспечивается за счет использования больших языковых моделей. В рассматриваемой архитектуре LLM не

используется как самостоятельный механизм распознавания эмоций, а выполняет функции высокоуровневой интерпретации результатов мультимодального анализа. На вход языковой модели передаются transcript коммуникации, временные метки, speaker timeline, эмоциональные признаки и behavioral metrics. Использование структурированного контекста позволяет существенно снизить вероятность генерации абстрактных или недостоверных выводов [6].

Большая языковая модель применяется для задач семантической суммаризации, интерпретации эмоциональной динамики, выявления рисков и генерации evidence-based рекомендаций. Формирование рекомендаций осуществляется на основе конкретных эпизодов коммуникации, что обеспечивает объяснимость результатов анализа и позволяет использовать систему не только как классификатор эмоций, но и как инструмент поддержки принятия решений в задачах управления коммуникациями.

Таким образом, предложенный подход к синтезу мультимодальных признаков позволяет повысить устойчивость и точность классификации эмоционального состояния субъектов коммуникации в условиях асинхронности каналов, шумов и неполноты данных. Использование адаптивной мультимодальной фузии, темпоральной синхронизации признаков и больших языковых моделей формирует основу для построения интеллектуальных систем анализа коммуникационного поведения нового поколения, ориентированных на комплексную обработку аудиовизуальной и семантической информации.

Литература

1. Барабанщиков В.А., Суворова Е.В. Выражение и восприятие мультимодальных эмоциональных состояний // Экспериментальная психология. 2023. Т. 16. № 4. С. 106–127.

2. Кушнир Д.А. Глубокое обучение в мультимодальных методах для распознавания эмоционального состояния диктора. Часть 1 // Известия высших учебных заведений. Приборостроение. 2023.
3. Лемаев В.И. Автоматическая классификация эмоций в речи: методы и данные // Информатика и автоматизация. 2024.
4. Ячная В.О. Современные технологии автоматического распознавания средств общения на основе визуальных данных // Научно-технические ведомости СПбПУ. 2023.
5. Поволоцкая А.А., и др. Запись и апробация набора речевых данных для распознавания негативных эмоций в речи // 2023.
6. Gandhi A., Adhvaryu K., Poria S., Cambria E., Hussain A. Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions // Information Fusion. 2023. Vol. 91. P. 424–444.
7. Wang Y., Song W., Tao W. et al. A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances // 2022.

Literature

1. Barabanshchikov V.A., Suvorova E.V. Expression and Perception of Multimodal Emotional States // Experimental Psychology. 2023. Vol. 16. No. 4. P. 106–127.
2. Kushnir D.A. Deep Learning in Multimodal Methods for Speaker Emotional State Recognition. Part 1 // News of Higher Educational Institutions. Instrument Engineering. 2023.
3. Lemaev V.I. Automatic Classification of Emotions in Speech: Methods and Data // Informatics and Automation. 2024.
4. Yachnaya V.O. Modern Technologies for Automatic Recognition of Communication Means Based on Visual Data // St. Petersburg Polytechnic University Journal. Engineering Science and Technology. 2023.
5. Povolotskaya A.A. et al. Recording and Validation of a Speech Dataset for Negative Emotion Recognition in Speech. 2023.

6. Gandhi A., Adhvaryu K., Poria S., Cambria E., Hussain A. Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions // Information Fusion. 2023. Vol. 91. P. 424–444.
7. Wang Y., Song W., Tao W. et al. A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances. 2022.