Мультимодальный анализ нового набора данных в системе обнаружения признаков обмана

А. А. Щеголева

Санкт-Петербургский государственный университет аэрокосмического приборостроения

aleksandrasheg@yandex.ru

***Аннотация*. Создана и обучена мультимодальная модель автоматизированного распознавания признаков лжи по видео- и аудиозаписям. Собран авторский набор данных, содержащий более 400 видеозаписей с размеченными метками «правда»/«ложь». Проведен сравнительный анализ трех подходов к слиянию модальностей: конкатенации, кросс-модальное внимание и кросс-модальное обучение с параметрической эффективностью. Анализ проведен в условиях приближенных к реальности – проверка на новых пользователях, съемка на устройства среднего качества и фоновый шум на записях.**

Ключевые слова: распознавание лжи, глубокое обучение, механизм внимания, мультимодальные модели

I. Введение

В современном мире потребность в надежных и быстрых инструментах верификации информации становится критически важной. Дистанционные инструменты скрининга находят все более широкое применение в таких областях, как HR-менеджмент (проверка кандидатов), банковский сектор (оценка рисков) и системы безопасности в аэропортах и на пограничном контроле. Традиционные методы, такие как классический полиграф, часто оказываются неудобными вследствие инвазивности, необходимости контактных датчиков и участия высококвалифицированных экспертов, что затрудняет их масштабирование для массового использования в реальном времени. Развитие удаленного формата работы и цифровизации сервисов обуславливает переход к бесконтактным методам детекции лжи, способным работать через стандартные каналы видеосвязи

В ряде современных работ уже представлены достаточно полные обзоры в этой области, включая анализ различных модальностей и архитектур машинного обучения. В исследованиях активно применяются [1–2]:

- визуальные признаки: мимика, микровыражения лица, единицы действия (Action Units), направление взгляда и жесты;
- акустические признаки: высота тона, спектральные характеристики (MFCC), паузы и заминки в речи;
- текстовые признаки: лингвистический анализ транскрипций, использование местоимений и когнитивная сложность высказываний;
- физиологические сигналы: ЭЭГ, КГР и термография, которые требуют специального оборудования [3].

Для обработки этих данных используются как классические модели (метод опорных векторов, случайный лес, KNN), так и глубокие нейронные сети (CNN, LSTM, BiLSTM, архитектуры на базе трансформеров) [4–5]. Установлено, что мультимодальный подход (объединение нескольких каналов данных) стабильно повышает точность классификации на 10–15% по сравнению с унимодальными методами

Основная проблема применения существующих подходов, что наблюдается культурная и языковая однородность: большинство наборов данных ориентировано на англоязычную или азиатские популяции (Индия, Китай), что затрудняет их перенос на другие этнические и культурные группы [6].

Наборы данных традиционно собираются тремя способами: в ходе реальных ситуаций с высокими ставками (например, судебные процессы), в игровых сценариях (игра «Оборотень» или «Мафия») [7] или через искусственно сгенерированные лабораторные опросы. При этом часто используется дорогостоящее оборудование, такое как высокоскоростные камеры или медицинские датчики

В данной работе решается проблема отсутствия репрезентативных данных для русскоязычной популяции. Основной акцент сделан на разработку системы, пригодной для массового использования, что предполагает работу с данными среднего качества: видеопотоком с фронтальной камеры смартфона и записью со стандартного микрофона телефона. В статье описывается процесс сбора собственного набора данных.

II. Описание нового набора данных

Основные характеристики используемого набора данных представлены в табл. I. Корпус сформирован при участии 30 уникальных пользователей в возрасте от 18 до 65 лет, каждый из которых предоставил по 20 видеозаписей, обеспечивая тем самым равномерный вклад каждого субъекта в итоговую выборку. Итоговый объём корпуса составил 463 видеозаписей. Участники были рекрутированы методом добровольного отклика.

ТАБЛИЦА I. ОСНОВНЫЕ ХАРАКТЕРИСТИКИ НАБОРА ДАННЫХ

Параметр	**Значение**
Общее количество видео	463 записи
Класс «правда»	239 записей (51,6%)
Класс «ложь»	224 записей (48,4%)
Уникальные пользователи	30 человек (каждый дал 20 видео)
Демография	Возраст 18-65 лет, 55% мужчины, 45% женщины
Длительность	Минимум 1 сек, максимум 60 сек, среднее ~10 сек

Запись материала осуществлялась в контролируемых, но приближённых к реальным условиях: участники располагались напротив камеры смартфона на расстоянии 40–60 см, освещение было естественным или стандартным офисным, фоновый шум не подавлялся. Подобные условия намеренно воспроизводят типичный сценарий дистанционного интервью, характерный для целевого применения системы.

Процедура прохождения интервью при сборе набора данных включала несколько этапов. На подготовительном этапе участник получал инструктаж об условиях эксперимента и подписывал форму информированного согласия на обработку аудиовизуальных персональных данных. Далее следовала разминочная серия нейтральных вопросов (имя, род занятий, хобби), позволявшая участнику адаптироваться к условиям записи и установить базовую линию поведения. Основная часть интервью включала блоки правдивых и ложных ответов, по 5–7 вопросов в каждом. В блоке правдивых ответов участник отвечал на вопросы в соответствии с реальными фактами своей биографии. В блоке ложных ответов — намеренно сообщал заведомо ложные сведения по заранее согласованному сценарию. Порядок предъявления блоков варьировался между участниками для контроля эффекта порядка. Каждый ответ фиксировался в виде отдельного видеофрагмента продолжительностью от 5 до 20 секунд и сопровождался метаданными: идентификатором участника, меткой класса (правда / ложь), номером вопроса.

Представленный набор данных характеризуется высокой плотностью записей на одного субъекта и сбалансированным распределением целевых классов, что критически важно для минимизации систематических ошибок при обучении мультимодальных моделей. Общий объем выборки составляет 463 видеозаписи, полученных от 30 участников, что в среднем соответствует 15,4 записи на одного человека. Изначальный протокол сбора данных предполагал наличие 20 видеофрагментов для каждого субъекта, однако итоговое число отражает результаты строгого контроля качества, в ходе которого фрагменты с техническими дефектами или недостаточной четкостью артикуляции были исключены.

Высокая репрезентативность индивидуальных паттернов поведения позволяет нейронной сети эффективно разделять субъект-специфичные признаки (уникальная мимика, тембр голоса, манера речи) и универсальные инварианты децепции. Это значительно снижает риск переобучения на конкретных личностях (identity bias) и способствует повышению обобщающей способности алгоритма при работе с новыми, ранее не встречавшимися дикторами.

Важным аспектом является внутренняя сбалансированность классов. Отношение количества ложных сообщений к истинным составляет 224/239= 0,937, что свидетельствует о практически идеальном паритете между целевыми метками. В контексте машинного обучения такой коэффициент близости к единице исключает необходимость применения специфических методов регуляризации, таких как взвешенные функции потерь или методы искусственного увеличения минорного класса (oversampling). Прямым следствием этого баланса является получение несмещенных оценок точности и полноты модели непосредственно в процессе обучения, что гарантирует стабильность градиентов и корректную настройку разделяющей гиперплоскости в латентном пространстве признаков.

Соблюдение условия субъектной независимости выборок являлось принципиальным требованием при формировании разбиений. Записи одного и того же пользователя включались строго в одно из разбиений — обучающее или валидационное, — что исключало возможность утечки субъект-специфичных признаков (индивидуальных паттернов мимики, тембра голоса, манеры речи) из обучающего подмножества в подмножество, предназначенное для валидации (рис. 1). Данное условие обеспечивает корректную оценку способности модели к обобщению на новых, ранее не встречавшихся пользователей, что критически важно для заявленного сценария применения системы.

Рис. 1. Схема распределения образцов на выборки

III. Условия численного моделирования стратегий слияния модальностей

В качестве входных данных используются векторы аудио- H^a и видеозаписи H^v, сформированные в рамках оригинального набора данных.

Экспериментальная часть исследования направлена на сравнительный анализ различных стратегий слияния модальностей, включая конкатенацию признаков (Concatenation of Features, CF):

$$Z^{CF} = \left[H^v \,||\, H^a \right]; \qquad (1)$$

кросс-модальное внимание (Cross-Modal Attention, CMA):

$$Z^{CMA} = soft\max\left(\frac{H^v \left(H^a\right)^T}{\sqrt{d}} \right) H^a; \qquad (2)$$

d – размерность векторов H^v и H^a.

CMA представляет собой математическую операцию, предназначенную для установления динамических зависимостей между двумя разнородными информационными потоками в едином латентном пространстве. В отличие от стандартного механизма самовнимания, где векторы запросов, ключей и значений извлекаются из одной последовательности, данная архитектура использует асимметричный подход к обработке данных. Применение функции мягкого максимума (Softmax) к нормализованной матрице сходства позволяет интерпретировать результат как

вероятностное распределение внимания, где каждый элемент отражает степень релевантности конкретного временного шага или признака второй модальности относительно текущего состояния первой.

CMA формирует обогащенное представление данных, которое сохраняет исходную структуру целевой модальности, но интегрирует в неё контекстуально важную информацию из вспомогательного источника.

Методика PECL [8]:

$$Z^{PECL} = f_\theta\left(H^a, H^v\right), \tag{3}$$

где f_θ – параметризованная модель с минимальным числом обучаемых параметров.

Основное отличие PECL подхода от CMA заключается в добавлении адаптеров, которые при обучении модели затрагивают меньшее количество весов модели. Вместо полного дообучения весов трансформера, внедряются легковесные блоки-адаптеры. Основный идея - захват временных зависимостей через одномерные свертки (1D-CNN) внутри стандартных слоев трансформера.

PECL переносит сложность обучения с огромных матриц весов трансформера $W \in \mathbf{R}^{D\times D}$ на малые матрицы адаптеров $W_{PECL} \in \mathbf{R}^{D\times d}$ и векторы внимания. Это позволяет адаптировать предобученные модели к задаче детектирования лжи, обучая менее 5–10% от общего числа параметров.

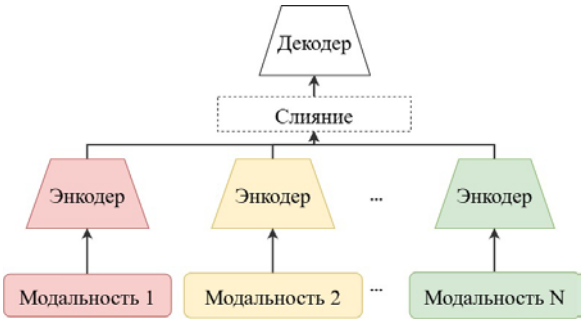

Рис. 2. Структура операции слияния признаков

С учетом различной длительности видеопоследовательностей применяется механизм маскирования, позволяющий выравнивать длины последовательностей при формировании батчей (batch) (пакетов обучающей выборки, обрабатываемых за один шаг обучения) и исключать влияние добавленных элементов на процесс обучения. Если T_i - длина i-го видео, $T_{max} = \max_i T_i$, то маска для i-го элемента батча определяется по формуле:

$$m_i^{(t)} = \begin{cases} 1, & t \le T_i, \\ 0, & t > T_i. \end{cases} \tag{4}$$

Для повышения устойчивости моделей используются методы аугментации данных, включающие пространственные и цветовые преобразования для видеоданных, а также временное маскирование для аудиосигналов.

Извлечение признаков осуществляется независимо для каждой модальности с использованием предобученных моделей, после чего полученные представления проецируются в общее латентное пространство. На заключительном этапе применяются различные стратегии слияния, формирующие единое представление, поступающее на вход классификатора.

Обучение модели осуществляется в сквозном режиме с использованием бинарной кросс-энтропийной функции потерь:

$$\mathcal{L}_{BCE} = -\frac{1}{\sum_i m_i}\sum_{i=1}^{N} m_i\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{5}$$

где $\hat{y}_i = \sigma(Z_i + b)$, $\sigma(Z_i) = 1/\left(1+e^{-Z_i}\right)$ - сигмоидальная функция, Z_i - линейная свертка входных переменных и обучаемых весов классификатора для i-го образца; N – размер батча (batch size, число объектов, обрабатываемых за одну итерацию), $y_i \in \{0,1\}$ - истинная бинарная метка i-го образца, $\hat{y}_i \in \{0,1\}$ - предсказанная вероятность положительного класса для i-го образца, $b \in \mathbb{R}$ - смещение классификатора.

IV. Обсуждение результатов моделирования

Из рис. 3 следует, что наихудшие показатели демонстрирует метод конкатенации признаков. Несмотря на относительную простоту реализации, данный подход характеризуется ограниченной способностью учитывать сложные зависимости между модальностями, что отражается в значениях accuracy и F1-меры.

Разброс значений метрик (в частности, полноты) указывает на нестабильность модели при изменении выборки, что может свидетельствовать о чувствительности к шуму и несбалансированности данных.

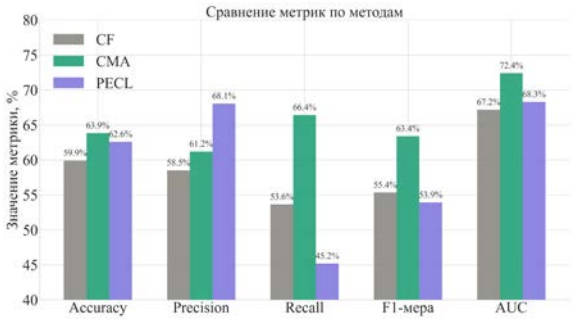

Рис. 3. Сравнение всех доступных метрик по трём методам

Использование кросс-модального механизма внимания приводит к заметному улучшению большинства показателей качества. Так, наблюдается рост accuracy и F1-меры, а также существенное увеличение recall, что свидетельствует о способности модели более полно выявлять целевой класс (табл. II).

ТАБЛИЦА II. Основные характеристики набора данных

Тип слияния	Accuracy	Precision	Recall	F1-мера	AUC
CF	59,91 ± 7,17	58,52 ± 7,97	53,65 ± 15,57	55,35 ± 1,082	67,18 ± 7,35
CMA	63,85 ± 2,24	61,21 ± 2,68	66,42 ± 8,23	63,40 ± 3,70	72,41 ± 3,38
PECL	62,59 ± 2,64	68,06 ± 5,39	45,20 ± 5,59	53,94 ± 4,17	68,30 ± 5,27

Повышение значения AUC дополнительно подтверждает улучшение разделяющей способности модели.

Низкие значения стандартного отклонения для точности указывают на более стабильное поведение по сравнению с конкатенацией.

Результаты подтверждают эффективность CMA, обеспечивающих адаптивное взвешивание признаков различных модальностей (рис. 4).

Метод PECL демонстрирует смешанные результаты:

- имеет место наивысшее значение precision, что указывает на снижение числа ложноположительных срабатываний и более «консервативный» характер классификации;
- наблюдается значительное снижение recall, что свидетельствует о пропуске значительной части положительных примеров.

Это приводит к относительно невысокому значению F1-меры, сопоставимому с результатами конкатенации. При этом значение AUC остается на среднем уровне, что указывает на приемлемую, но не максимальную способность модели к разделению классов.

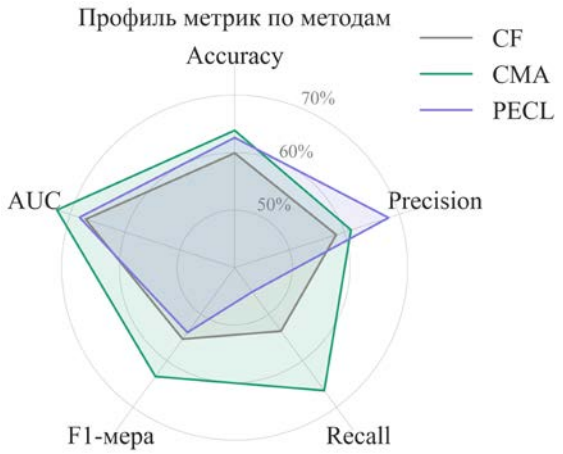

Рис. 4. Профиль метрик по методам слияния

Соотношение Precision/Recall для PECL составляет 68,06/45,20 = 1,51, что свидетельствует о выраженном «пессимистическом» смещении модели: она склонна скорее не предъявить обвинение, чем предъявить ложное. В отличие от CMA, где соотношение Precision/Recall = 61,21/66,42 = 0,92 (лёгкий «оптимистический» уклон), PECL принципиально иначе управляет балансом ошибок первого и второго рода, что имеет принципиальное значение при выборе модели для конкретного прикладного сценария (например, банковский скрининг).

Таким образом, наибольшая сбалансированность между recall и precision достигается при использовании кросс-модального внимания, которое обеспечивает наиболее высокие значения интегральных метрик качества.

Однако подход PECL может быть предпочтителен в сценариях, где критически важно минимизировать ложноположительные ошибки.

Метод конкатенации, несмотря на простоту, уступает более современным подходам и может рассматриваться лишь как базовый ориентир при построении мультимодальных моделей.

V. Заключение

В докладе представлены новый оригинальный обучающий набор для распознавания целевого признака (ложь) и результаты сравнения типов слияния модальностей на основе мультимодальной системы автоматизированного обнаружения признаков обмана [9], ориентированной на работу в условиях, приближённых к реальным: видеопоток с фронтальной камеры смартфона и запись со стандартного микрофона без специализированного оборудования.

Представлен авторский набор данных, включающий 463 размеченные видеозаписи 30 участников русскоязычной популяции с соблюдением принципа субъектной независимости при формировании обучающего и валидационного разбиений. Данный корпус восполняет существенный пробел в области детекции лжи — отсутствие репрезентативных данных для русскоязычной аудитории.

Проведённый сравнительный анализ трёх стратегий слияния модальностей показал, что метод кросс-модального внимания (CMA) обеспечивает наилучший баланс между точностью и полнотой классификации, достигая значений accuracy 63,85% и AUC 72,41%.

Метод PECL демонстрирует наибольшую точность (precision 68,06%) при сниженном recall, что делает его предпочтительным в сценариях с высокой ценой ложноположительных ошибок.

Конкатенация признаков уступает обоим подходам и может рассматриваться лишь как базовый ориентир.

Полученные результаты подтверждают перспективность мультимодального подхода для бесконтактной дистанционной детекции лжи.

В качестве направлений дальнейших исследований автор видит расширение корпуса данных, привлечение более широкой демографической выборки, апробацию созданного программного комплекса на прикладных задачах по выявлению значимого признака, а также исследование более глубоких архитектур слияния модальностей и методов адаптации к индивидуальным особенностям пользователей.

Список литературы

[1] Rahayu Y.D., Fatichah C., Yuniarti A. and Rahayu Y.P. Advancements and Challenges in Video-Based Deception Detection: A Systematic Literature Review of Datasets, Modalities, and Methods. *IEEE Access*. 2025. Vol. 13. Pp. 28098-28122.

[2] D'Ulizia A., D'Andrea A., Grifoni P., Ferri F. Analysis, Evaluation, and Future Directions on Multimodal Deception Detection. *Technologies*. 2024. Vol. 12. P. 71.

[3] Joshi G., Tasgaonkar V., Despande A. et al. Multimodal machine learning for deception detection using behavioral and physiological data. *Scientific Reports*. 2025. Vol. 15. P. 92399.

[4] Prome S.A., Ragavan N.A., Islam MR. et al. Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review. *Natural Language Processing Journal*. 2024. Vol. 6. P. 100057.

[5] Talaat F.M. Explainable Enhanced Recurrent Neural Network for lie detection using voice stress analysis. *Multimedia Tools and Applications*. 2024. Vol. 83. P. 32277–32299.

[6] Joshi G., Tasgaonkar V., Despande A. et al. Multimodal machine learning for deception detection using behavioral and physiological data. *Scientific Reports*. 2025. Vol. 15. P. 92399.

[7] Yang C., You X., Xie X. et al. Development of a Chinese werewolf deception database. *Frontiers in Psychology*. 2023. Vol. 13. P. 1047427.

[8] Guo X., Selvaraj N.M., Yu Z. et al. Audio-Visual Deception Detection: DOLOS Dataset and Parameter-Efficient Crossmodal Learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. P. 22078–22088.

[9] Свидетельство о государственной регистрации программы для ЭВМ № 2025692560. Программный комплекс для распознавания признаков лжи в видеозаписях/ Щеголева А.А. (RU). – Заявка № 2025690840; Дата поступления 11 ноября 2025; Зарегистрировано 21 ноября 2025.