Применение глубоких аудиопризнаков для распознавания лжи в речи

В. А. Матвеев

Санкт-Петербургский государственный университет аэрокосмического приборостроения

Email: matveev856@gmail.com

А. А. Щеголева

Санкт-Петербургский государственный университет аэрокосмического приборостроения

***Аннотация*. В рамках работы предложен метод автоматизированного распознавания лжи в речи, базирующийся на анализе глубоких аудиопризнаков. Основой исследования стал авторский датасет из более 400 верифицированных записей. Проведен сравнительный анализ эффективности обработки речевых сигналов: исследован результат применения классических Mel-спектрограмм в качестве входных данных для сверточных нейронных сетей, а также применение современных архитектур на основе трансформеров и предобученных векторных представлений данных (аудио-эмбеддингов). Особое внимание уделено обобщающей способности системы: модель прошла тестирование в условиях реальной акустической среды (фоновые шумы, устройства среднего качества) с применением кросс-пользовательской валидации, что подтвердило преимущество глубоких методов извлечения признаков над традиционным спектральным анализом при работе с нелингвистическими маркерами обмана.**

Ключевые слова: распознавание лжи, анализ речевых сигналов, глубокое обучение, паралингвистика, акустические признак, классификация аудиоданных, машинное обучение

I. Введение

Автоматизированное распознавание лжи представляет собой одну из наиболее актуальных и методологически сложных задач в области национальной безопасности, юриспруденции и психологии [1, 2]. Ложь является сложным когнитивным процессом, сопровождающимся разнообразными поведенческими и физиологическими маркерами, проявляющимися в том числе на уровне речевого сигнала [3]. Традиционные инструменты верификации правдивости показаний, в частности, полиграф, подвергаются систематической критике ввиду их инвазивности, зависимости от субъективных экспертных оценок и уязвимости к применению контрмер [4, 5]. Это обусловило значительный рост интереса исследовательского сообщества к неинвазивным автоматизированным методам детекции лжи на основе технологий искусственного интеллекта [6-8].

На начальном этапе развития систем автоматического распознавания лжи по голосу для извлечения акустических признаков преимущественно применялись классические методы [9]: кепстральные коэффициенты на мел-шкале (Mel-Frequency Cepstral Coefficients , MFCC), основная частота тона (F0), а также характеристики энергии сигнала и просодические параметры. С развитием методик глубокого обучения [10] (Deep Learning) исследователи перешли к использованию сверточных нейронных сетей (CNN) для анализа визуальных представлений звука — Mel-спектрограмм. Наиболее современные подходы предполагают применение архитектур трансформера и предобученных аудио-эмбеддингов (например, Wav2Vec 2.0), способных извлекать высокоуровневые скрытые представления, недоступные для традиционного спектрального анализа.

Ключевой нерешённой проблемой остаётся обобщающая способность разрабатываемых систем в условиях реальной эксплуатации. Большинство существующих моделей обучаются на лабораторных датасетах с контролируемым качеством записи, что приводит к существенному снижению точности при столкновении с фоновыми шумами, использованием устройств среднего качества или индивидуальными особенностями речи новых пользователей. Преодоление «доменного сдвига» и обеспечение инвариантности к акустическим условиям являются критическими условиями для практического внедрения систем детекции лжи.

Целью настоящей работы является разработка и верификация метода автоматизированного распознавания лжи, основанного на глубоком анализе аудиопризнаков. В рамках исследования проводится сравнительный анализ эффективности классических Mel-спектрограмм в связке с CNN и современных трансформерных архитектур. Работа опирается на верифицированный датасет из 962 записей и фокусируется на оценке устойчивости предложенных алгоритмов к условиям реальной акустической среды с применением строгой кросс-пользовательской валидации.

II. Подготовка и обработка данных

A. Характеристики исходного набора данных (рис. 1)

Для проведения исследования использовался верифицированный датасет, включающий аудиозаписи 32 уникальных пользователей. Исходные материалы представляли собой видеофайлы разрешением 1280×720, из которых были извлечены моноаудиодорожки с частотой дискретизации 16 кГц. Общая длительность аудиоматериала составила 2 часа 14 минут 15 секунд. Распределение классов является достаточно сбалансированным: 52,9% записей отнесены к категории «Ложь» (класс 0), 47,1% — к категории «Правда» (класс 1) (рис. 1). Записи производились на основе фронтальных камер мобильных телефонов со встроенными микрофонами, что ограничивает

воспроизводимый частотный диапазон верхним порогом около 15 кГц и приближает условия эксперимента к реальной акустической среде.

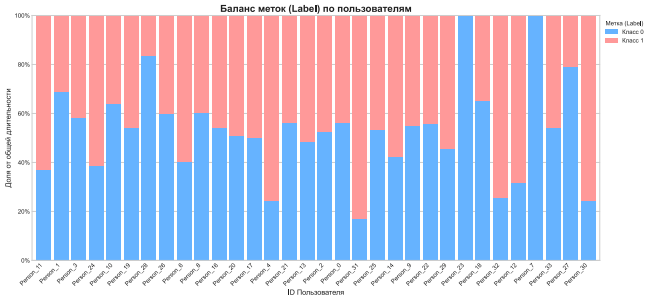

Рис. 1. Распределение по пользователям внутри датасета

B. Предварительная обработка аудиосигнала

Конвейер предобработки включает три последовательных этапа (рис. 2).

- **Нормализация амплитуды:** Все сигналы приводились к единому диапазону амплитуд [-1.0; 1.0], что исключает влияние абсолютной громкости голоса на принимаемые классификатором решения.
- **Обрезка тишины (Trim):** Паузы в начале и конце каждой записи удалялись с целью исключения неинформативных сегментов, занимавших значительную долю суммарного времени датасета.
- **Сегментация (Sliding Window):** Очищенный сигнал разбивался на перекрывающиеся фрагменты (чанки) длительностью 2 секунды с шагом 1 секунда (overlap 50%). Данный подход позволяет кратно увеличить объём обучающей выборки, а также зафиксировать кратковременные всплески эмоционального стресса, характерные для состояния обмана.

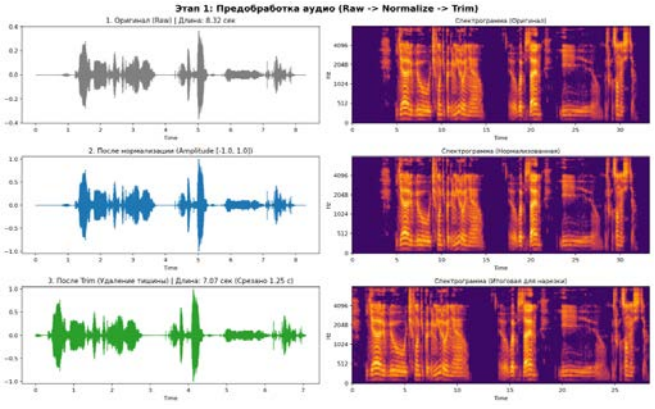

Рис. 2. Предобработка аудио

III. Архитектура системы

В качестве базовой (baseline) архитектуры была выбрана модель «сиамских близнецов» (Siamese Network) с двухэтапной процедурой обучения. Выбор данной архитектуры обусловлен её способностью обучаться на метрическом пространстве представлений, что позволяет эффективно сравнивать пары примеров без необходимости прямой классификации на первом этапе.

Энкодер. Общий для обоих входов энкодер f_θ состоит из трёх последовательных свёрточных блоков, каждый из которых включает одномерную свёртку Conv1D, батч-нормализацию, функцию активации ReLU и слой пулинга. Выход свёрточной части выравнивается и подаётся на полносвязный слой, состоящий из линейного преобразования, активации ReLU, dropout и финального линейного слоя, проецирующего представление в пространство эмбеддингов размерности 128.

Этап 1: Метрическое обучение энкодера. На первом этапе энкодер f_θ обучается в сиамской конфигурации: одна и та же сеть применяется к обоим элементам входной пары (запись правды и записи лжи), порождая эмбеддинги e_truth и e_lie. Обучение производится посредством минимизации контрастивной функции потерь (Contrastive Loss), в качестве показателя используется евклидово расстояния.:

A. Contrastive loss

$$L(z_i, z_j, y) = (1-y) * \frac{1}{2}\|z_i - z_j\|_2^2 + y * \frac{1}{2}[max(0, m - \|z_i - z_j\|^2]^2$$

z_i, z_j — эмбеддинги входной пары, $y \in \{0,1\}$ — метка схожести пары (0 — схожие, 1 — несхожие), m — параметр отступа (margin).

Данная функция потерь побуждает энкодер сближать эмбеддинги записей одного класса и разводить эмбеддинги разных классов на расстояние не менее m. По завершении первого этапа веса энкодера замораживаются.

Этап 2: Обучение классификатора. На втором этапе поверх замороженного энкодера обучается классификатор. В качестве входного признака используется поэлементная абсолютная разность эмбеддингов |e_1 − e_2|, которая отражает расстояние между парой в метрическом пространстве. Классификатор состоит из линейного слоя с 64 нейронами, активации ReLU, dropout и выходного линейного слоя с функцией активации Sigmoid, возвращающего вероятность принадлежности к классу «ложь». Обучение классификатора производится посредством минимизации бинарной кросс-энтропии (BCELoss). Разделение этапов обучения позволяет независимо оценить качество метрических представлений и классификационной головы, а также упрощает интерпретацию результатов.

Система тестировалась с двумя видами входных представлений векторов-признаков (табл. 1):

ТАБЛИЦА I. МЕТРИКИ ДЛЯ СИАМСКОЙ МОДЕЛИ ДЛЯ РАЗНЫХ ВХОДНЫХ ПАРАМЕТРОВ

№	EMB	F1	ACC
1	MFCC	0.54	0.54
2	WAV2VEC 2.0	0.51	0.52

Полученные метрики соответствуют уровню точности, характерному для человека при решении аналогичной задачи — по данным ряда исследований, люди распознают ложь в среднем с точностью около 54% [1], что лишь незначительно превышает случайное угадывание. Таким образом, базовая модель воспроизводит «человеческий потолок» для данной

задачи, однако это одновременно указывает на значительный потенциал для улучшения: применение более специализированных архитектур и расширение обучающей выборки должны позволить системе существенно превзойти возможности человека-эксперта.

IV. Дальнейшие перспективы

Для повышения качества классификации и обобщающей способности системы планируется развитие исследования по следующим направлениям.

- **Расширение обучающей выборки.** Текущий датасет из 400+ записей является ограниченным для обучения глубоких моделей. В качестве методов увеличения объёма данных рассматриваются: аугментация аудиоданных (добавление фонового шума, изменение темпа и высоты тона, применение реверберации)
- **Архитектура с градиентным инверсионным слоем (GRL).** Для повышения инвариантности модели к индивидуальным особенностям голоса конкретных пользователей планируется внедрение слоя реверсирования градиента (Gradient Reversal Layer, GRL) в рамках архитектуры состязательного доменного обучения (Domain-Adversarial Neural Network, DANN). Данный подход позволяет явно разделить извлечение признаков лжи, инвариантных к диктору, от признаков, специфичных для конкретного пользователя, что напрямую решает проблему «доменного сдвига» при кросс-пользовательской валидации.

Список литературы

[1] Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review.Shanjita Akter Prome a, Neethiahnanthan Ari Ragavan b, Md Rafiqul Islam c, David Asirvatham a, Anasuya Jegathevi Jegathesan // Natural Language Processing Journal, Volume 6, March 2024, 100057.

[2] Shanjita Akter Prome, Neethiahnanthan Ari Ragavan, Md Rafiqul Islam, David Asirvatham, Anasuya Jegathevi Jegathesan, Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review // Natural Language Processing Journal, Volume 6, 2024, 100057, ISSN 2949-7191, https://doi.org/10.1016/j.nlp.2024.100057.

[3] Analysis, Evaluation, and Future Directions on Multimodal Deception Detection. Arianna D'Ulizia, Alessia D'Andrea *, Patrizia Grifoni and Fernando Ferri. DOI:10.3390/technologies12050071

[4] Deception Detection from Linguistic and Physiological Data Streams Using Bimodal Convolutional Neural Networks. Panfeng Li e.t.c https://doi.org/10.48550/arXiv.2311.10944

[5] Каширина Е.И. Современные технологии дистанционного детектирования лжи: обзор готовых решений и перспективы развития / Е. И. Каширина, Д. А. Болтышев, Д. В. Мочалов // Электронный сетевой политематический журнал "Научные труды КубГТУ". 2023. № 6. С. 90-98. – EDN LCXELE.

[6] SVC 2025: the First Multimodal Deception Detection Challenge. Xun Lin e.t.c

[7] Y. D. Rahayu, C. Fatichah, A. Yuniarti and Y. P. Rahayu, "Advancements and Challenges in Video-Based Deception Detection: A Systematic Literature Review of Datasets, Modalities, and Methods," in IEEE Access, vol. 13, pp. 28098-28122, 2025, doi: 10.1109/ACCESS.2025.3533545

[8] Chen, H., Chai, Z., Dogru, O., Jiang, B., & Huang, B. (2022). Data-Driven Designs of Fault Detection Systems via Neural Network-Aided Learning. // IEEE Transactions on Neural Networks and Learning Systems, 33(10), 5694–5705. https://doi.org/10.1109/tnnls.2021.3071292

[9] Kusumawati, Dewi et al. "Performance Analysis of Feature Mel Frequency Cepstral Coefficient and Short Time Fourier Transform Input for Lie Detection using Convolutional Neural Network." // JOIV: International Journal on Informatics Visualization (2024): n. pag.

[10] Talaat, Fatma M. "Explainable Enhanced Recurrent Neural Network for lie detection using voice stress analysis." // Multimedia Tools and Applications 83 (2023): 32277-32299.