Модуль выявления применения AI-VC в аудиосообщениях

Н. Ю. Мирошников, А. Д. Шульженко

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

nicolasmiroshnikov@gmail.com

Аннотация. **В работе предложена концепция модуля для автоматического выявления применения AI-VC в голосовых сообщениях в подсистеме «мессенджер» ИС «смартфон». Выполнен обзор ключевых научных статей, подтверждающих актуальность атак с использованием AI-VC. Проведен анализ, по результатам которого выделены наиболее характерные параметрические признаки, используемые для детекции AI-VC, а также предложено агрегирование двух подходов к внедрению предложенного модуля. Проведена оценка экономической целесообразности внедрения модуля.**

Ключевые слова: преобразование голоса; AI-VC; дипфейк; противодействие вишингу; детектирование артефактов AI-VC

I. Введение

Развитие средств преобразования голоса (voice conversion, VC) и появление VC-средств, использующих искусственный интеллект (AI-VC), привели к появлению угроз, связанных с подменой голосовой идентичности. Такие угрозы актуальны в различных программных подсистемах смартфона, а особенно – в мессенджерах.

Первым крупный случай мошенничества с использованием AI-VC произошёл в 2019 году, когда работник английской энергетической фирмы перевёл 243.000 долларов США на счёт мошенника, представившегося его начальником [1]. При этом в исследовании [2], проведённом в 2024 году, было выявлено, что человек не способен достоверно распознавать голосовые дипфейки, что говорит о человеческой уязвимости и необходимости внедрения автоматики в системы, обрабатывающие голосовые записи. В рамках работы [3] были разработаны две модели, способные к выявлению дипфейк-записей на базовом уровне ASVspoof 2019, но обладающие низкой способностью к выявлению преобразованной речи в режиме живого времени в звонках на платформе Teams. Указанные работы подтверждают актуальность угрозы атак с применением голосовых записей, созданных с помощью AI-VC.

Несмотря на то, что методы обнаружения факта преобразования голоса с помощью AI-VC, предлагаемые в современных работах, показывают высокую эффективность в рамках исследовательских тестов, их интеграция на уровне подсистем смартфона остаётся малоисследованной. Внедрение модуля детектирования дипфейк-голосов в мессенджеры может существенно повысить устойчивость к атакам с использованием AI-VC. При выявлении признаков синтетической речи мессенджер, по предлагаемой концепции, инициирует пуш-уведомление через API операционной системы, тем самым оказывая управляющее воздействие на работу смартфона и повышая его устойчивость к голосовым атакам. Настоящая работа посвящена описанию принципов работы предлагаемого модуля на основе двух наиболее значимых групп признаков, позволяющих определить преобразование голоса с помощью AI-VC, и интеграции его в рамках аппаратно-программной ИС «смартфон».

II. Признаки применения AI-VC к голосовой записи

Средства преобразования голоса на основе AI имеют различные архитектурные особенности, но общий принцип работы заключается в извлечении параметрических признаков из аудиозаписей речи исходного говорящего, преобразовании их в признаки целевого говорящего и последующем восстановлении полученных признаков в аудиоформат с помощью нейро-вокодера. Искажения, наблюдаемые в записях с преобразованными голосами, обусловлены действиями, производимыми на перечисленных этапах. Так, в работе [4] предлагается метод определения преобразования голоса по артефактам, вносимым нейро-вокодером. Обнаружение наличия подобных артефактов возможно благодаря анализу различных параметрических признаков аудиосигнала с речью.

В табл. 1 приведены основные параметрические признаки, используемые в различных методах обнаружения преобразования голоса и методы, использующие их. Как можно заметить, наиболее часто используемым из них являются CQCC, то есть кепстральные коэффициенты на основе постоянного преобразования Q (алгоритм преобразования ряда данных в частотную область). Так, 26 из 48 участников ASVspoof 2019 использовали CQCC в своих системах верификации говорящего [5]. Связано это с тем, что CQCC обладают высокой спектральной точностью в области низких частот и высоким временным разрешением в области верхних. Эти преимущества позволяют надёжно выявлять артефакты AI-VC, которые традиционные кепстральные признаки (MFCC, LFCC) могут пропустить.

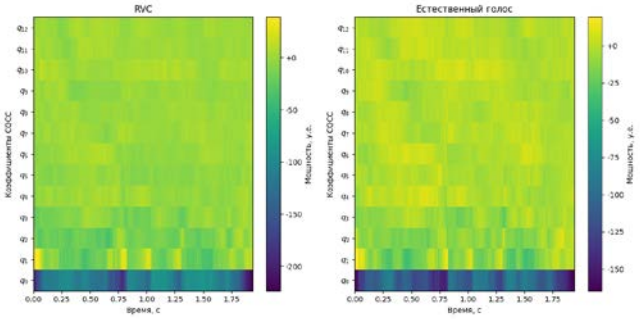

Рис. 1. Сравнение CQCC естественного голоса и голоса, преобразованного AI-VC (RVC)

ТАБЛИЦА I. ПАРАМЕТРИЧЕСКИЕ ПРИЗНАКИ ПРЕОБРАЗОВАНИЯ РЕЧИ, ИСПОЛЬЗУЕМЫЕ В РАЗЛИЧНЫХ МЕТОДАХ ОБНАРУЖЕНИЯ AI-VC

Методы	**Параметрические признаки**						
	LPC-residual	***MFCC***	***LFCC***	***CQCC***	***ZCR***	***F_0 (pitch, jitter, shimmer)***	***Spectral flux***
MFCC-ResNet		+					
CQCC-ResNet				+			
LSTM/TCN/CRNN-Spoof			+	+			
DeepSonar					+		+
Audio deepfake detection Integrating CNN+BiLSTM		+		+		+	
SSAD 2020			+	+			+
SASD 2022				+	+		
LP-residual features (RMFCC, RCQCC, RPCQCC)	+	+		+	+		

Несмотря на сказанное выше, мел-кепстральные коэффициенты (MFCC) также являются важным признаком, используемым в определении применения AI-VC к аудиозаписям. MFCC строятся на мел-частотной шкале, имитирующей неравномерную чувствительность человеческого слуха к разным частотам, что позволяет захватить основные характеристики спектральной оболочки речи, которые AI-VC системы часто воспроизводит менее естественно. MFCC являются одним из наиболее часто используемых признаков в ASVspoof на ряду с CQCC, потому что задают «базовый уровень обнаружения», от которого отталкиваются многие методики при добавлении более сложных признаков.

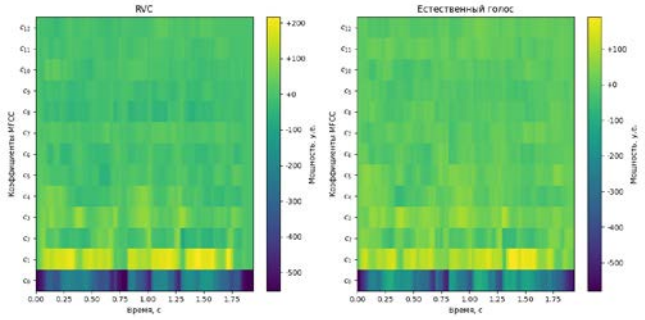

Рис. 2. Сравнение MFCC естественного голоса и голоса, преобразованного AI-VC (RVC)

Ещё одним популярным признаком является ZCR (Zero Crossing Rate). Обусловлено это тем, что нейровокодеры (один из краеугольных этапов работы AI-VC систем) нередко собирают волновую форму «по частям», что приводит к мелким колебаниям фазы и всплескам высокочастотного шума. ZCR измеряет частоту пересечений нуля во временной области, поэтому сразу реагирует на подобные артефакты, которые в натуральной речи проявляются значительно менее выраженно. Также ZCR имеет сравнительно низкую вычислительную сложность по сравнению с кепстральными и спектральными признаками, что может быть крайне актуально при edge-вычислениях на мобильных устройствах. Таким образом, ZCR выступает легковесным дополнением к тяжёлым кепстральным признакам, таким как CQCC и MFCC, позволяющим ещё точнее выявлять использование AI-VC в аудиозаписях.

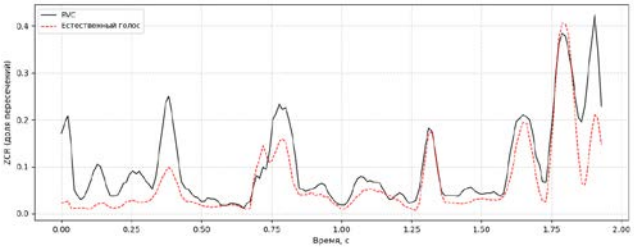

Рис. 3. Сравнение ZCR естественного голоса и голоса, преобразованного AI-VC (RVC)

III. ПРИНЦИП РАБОТЫ ПРЕДЛАГАЕМОГО МОДУЛЯ

Принцип работы предлагаемого модуля совместно с ИС «смартфон» и её подсистемой «мессенджер» заключается в следующем:

- пользователь А отправляет аудиосообщение пользователю Б в мессенджере;
- мессенджер передаёт аудиоданные в модуль детекции AI-VC;
- модуль последовательно выполняет предобработку (извлечение признаков) и классификацию (определение метки класса «AI-VC» / «Естественная речь»);
- мессенджер получает ответ от модуля в виде метки класса. В случае, если выставлена метка «AI-VC», мессенджер инициирует пуш-уведомление с предупреждением «Обнаружено аудиосообщение, сгенерированное AI» через API ОС смартфона пользователя Б.

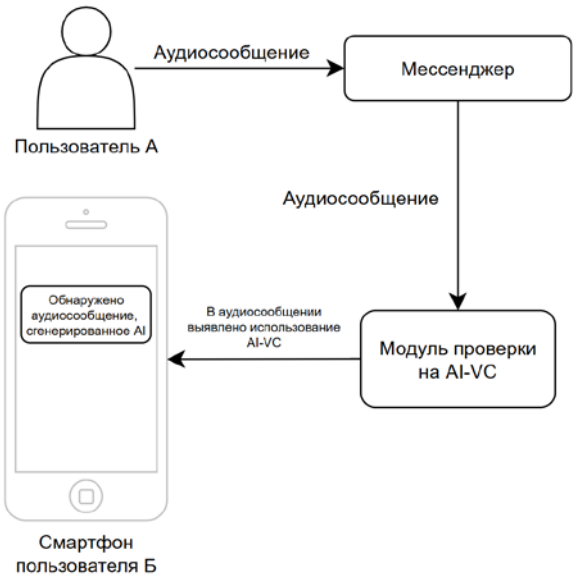

Рис. 4. Принцип работы предлагаемого модуля

Описанный принцип работы не препятствует использованию средств AI-VC напрямую, а направлен на информирование пользователей о возможности атак с использованием таких средств. Таким образом, модуль не мешает коммуникации пользователей в случае, если будет допущена ошибка первого рода, как если бы от его решения зависел сам факт доставки сообщения, а лишь стремится повысить бдительность пользователей.

С точки зрения развёртывания подобного модуля существует два принципиально отличающихся подхода: централизованный, при котором все сообщения пересылаются на серверы для обработки, и распределённый, при котором инференс (предобработка и классификация) для каждого из аудиосообщений осуществляется непосредственно на смартфоне отправителя, то есть используются edge-вычисления.

Согласно исследованию [6] извлечение первых 13 MFCC и ZCR из 1 с аудио с помощью Python на Apple MacBook Pro (2.3 GHz Intel Core i5) занимает ~20 мс, где ~16 мс уходят на вычисление MFCC. Если предположить, что вычисление CQCC занимает сопоставимое время в силу схожести производимых математических операций, то общее время предобработки для трёх самых популярных признаков составит ~50 мс (с запасом). При этом время классификации 4 с аудио моделями детекции AI-VC в исследовании [7] составило ~27–152 мс на схожем с предыдущим примером CPU (2 GHz Quad–Core Intel Core i5), что при пересчёте на 1 с аудио даёт время классификации, равное ~7–38 мс. Таким образом, суммарное время инференса на сервере под управлением CPU архитектуры x86 составит ~90–100 мс. Современные смартфонные CPU архитектуры ARM по производительности на ядро уступают процессорам архитектуры x86 в ~2–3 раза согласно независимым бенчмаркам [8]. Исходя из этого факта, можно утверждать, что время инференса на смартфоне под управлением CPU архитектуры ARM составит ~200–300 мс.

Согласно отчёту компании Meta [9], пользователи их мессенджера «WhatsApp» ежедневно отправляют по 7 млрд. голосовых сообщений. Это означает, что при равномерной нагрузке в секунду отправляются 81 тыс. аудиосообщений. Если выполнять все 81 000 запросов к классификатору на серверном CPU, то при среднем времени инференса в ~100 мс потребуется порядка 8100 ядер современных ускорителей, что равно 2025 процессорам Intel Core i5, 2,3 GHz, а суммарное энергопотребление такого кластера достигнет ~59 кВт в пике (81 тыс. запросов в секунду) при TDP одного процессора, равному ~29 Вт. Приведённые расчёты, несмотря на свою грубость, показывают, что централизованная реализация является дорогой и сложной в масштабировании. При этом, с учётом специфики применения описываемого модуля, даже потенциальная задержка в 0.5 с при инференсе на смартфоне не будет являться критичной, потому что модуль призван лишь уведомлять пользователя о потенциальной угрозе, а значит работать в параллели с подсистемой доставки сообщений, не препятствуя её работе.

Таким образом, хотя серверный вариант может обеспечить централизованный контроль и единообразие версий модуля, с учётом объёмов в миллиарды сообщений в день и высоких требований к вычислительным ресурсам и энергии, предпочтительнее внедрять модуль AI-VC-детекции непосредственно на смартфонах пользователей.

IV. Заключение

В результате выполнения работы была представлена концепция модуля выявления применения AI-VC в аудиосообщениях, предлагаемая к внедрению в ИС «смартфон». Для оценки экономической целесообразности внедрения подобной технологии были выделены наиболее часто используемые параметрические признаки, позволяющие выявить применение AI-VC и произведены расчёты затрат, необходимых для внедрения модуля, работающего на основе выявленных признаков, такие как оборудование (CPU) и электроэнергия.

V. Дальнейшие перспективы исследования

Дальнейшие исследования могут включать более детальную и точную экономическую оценку внедрения подобного модуля в ИС «смартфон» с опорой на конкретную модель классификатора и используемые в нём признаки, а также на оптимизированные средства извлечения признаков на этапе предобработки. Также, в связи с актуальностью упомянутых в начале доклада AI-VC атак, производимых посредством звонков, значительный интерес представляет разработка и оценка возможности интеграции модуля, подобного описанному, но способного работать в режиме живого времени.

Список литературы

[1] Jesse D. A Voice Deepfake Was Used To Scam A CEO Out Of $243,000 [Электронный ресурс] URL: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/ (дата обращения: 28.06.2025)

[2] Kamil M., Anton F., Milan Š., Daniel P., Karolína R., Petr H. Comprehensive multiparametric analysis of human deepfake speech recognition [Электронный ресурс] // EURASIP Journal on Image and Video Processing. 2024. Article 24. URL: https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-024-00641-4 (Дата обращения 29.06.2025)

[3] Jonat M., Rakin A., Sae F., Jagdish K., Huzaifa P., Agamjeet P., Sara A., Madhu R., Arjun P. Towards the Development of a Real-Time Deepfake Audio Detection System in Communication Platforms

[Электронный ресурс] URL: https://arxiv.org/abs/2403.11778 (Дата обращения: 30.06.2025)

[4] Chengzhe S., Shan J., Shuwei H., Siwei L. AI-Synthesized Voice Detection Using Neural Vocoder Artifacts [Электронный ресурс] URL: https://arxiv.org/abs/2304.13085 (Дата обращения: 30.06.2025)

[5] Hemlata T., Jose P., Andreas N., Nicholas E., Massimiliano T. An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification [Электронный ресурс] URL: https://arxiv.org/abs/2004.06422 (Дата обращения: 30.06.2025)

[6] Bruno S., Axel H., An B., Abdellah T. Evaluation of Classical Machine Learning Techniques towards Urban Sound Recognition on Embedded Systems [Электронный ресурс] URL: https://www.mdpi.com/2076-3417/9/18/3885 (Дата обращения: 01.07.2025)

[7] Piotr K., Marcin P., Piotr S. Towards Faster and More Accessible Audio DeepFake Detection [Электронный ресурс] URL: https://arxiv.org/abs/2210.06105 (Дата обращения: 01.07.2025)

[8] Сравнение производительность процессоров Intel Core i7-1165G7 и Rockchip ARMv8 Cortex-A76 4-Core [Электронный ресурс] URL: https://openbenchmarking.org/vs/Processor/Intel%2BCore%2Bi7-1165G7%2CRockchip%2BARMv8%2BCortex-A76%2B4-Core (Дата обращения: 01.07.2025)

[9] New Voice Message Features on WhatsApp [Электронный ресурс] URL: https://about.fb.com/news/2022/03/new-voice-message-features-on-whatsapp/ (Дата обращения:01.07.2025)