

Н.Ю. МИРОШНИКОВ, У.В. ТОКАРЕВА, А.Д. ШУЛЬЖЕНКО
**СПОСОБ ОПРЕДЕЛЕНИЯ АИ-ПРЕОБРАЗОВАННЫХ РЕЧЕВЫХ
АУДИОЗАПИСЕЙ НА ОСНОВЕ АКУСТИЧЕСКИХ
И СТАТИСТИЧЕСКИХ ПРИЗНАКОВ**

Мирошников Н.Ю., Токарева У.В., Шутьженко А.Д. Способ определения АИ-преобразованных речевых аудиозаписей на основе акустических и статистических признаков.

Аннотация. В работе предложен способ определения факта преобразования аудиосигнала с использованием интеллектуальных средств голосового синтеза (AI-based Voice Conversion, AI-VC) речевых аудиозаписей (например, с целью подмены автора или смыслового наполнения) на основе анализа акустических и статистических признаков аудиосигнала. Рассмотрены принципы работы современных систем преобразования голоса, сделан акцент на архитектуре преобразования голоса на основе поиска (Retrieval-based Voice Conversion, RVC). Показано, что преобразование представляет собой поэтапный процесс, в ходе которого происходит извлечение признаков целевого голоса из исходной речевой аудиозаписи, что эквивалентно векторизации бихевиаристических свойств говорящего, построение целевого шаблона – модели для синтеза голоса, формирование необходимого смыслового наполнения и синтез новой речевой аудиозаписи, имеющей акустически незначимое различие с целевой. Проведен обзор информативных признаков, чувствительных к такому преобразованию и принятых как артефакты синтеза: мел-частотные и линейно-частотные кепстральные коэффициенты (Mel-Frequency Cepstral Coefficients, MFCC, Linear-Frequency Cepstral Coefficients, LFCC), кепстральные коэффициенты на основе постоянного Q (Constant Q Cepstral Coefficients, CQCC), статистические свойства ошибки предсказания (Linear Predictive Coding, LPC), контур основной частоты и его вариативность (джиттер, шиммер), спектральный центронид и поток, скорость пересечения нуля. Разработана формальная модель бинарной классификации и программная реализация на основе гауссовых смесей (Gaussian Mixture Model, GMM). Экспериментальная проверка предложенного способа на наборе данных ASVspoof2019 показала точность классификации 93,9% (AUC = 0,939, EER = 13,52%). Это говорит о том, что предложенный способ может обеспечить защиту получателей речевых аудиозаписей от целевых вишинговых атак по типу спуфинга (мошенничества с подменой голосовой идентичности). Способ может быть реализован в виде программного модуля и интегрирован в мессенджеры или корпоративные коммуникационные платформы. Актуальность работы обусловлена отсутствием готовых продуктов, определяющих голосовые подделки на основе результатов спектрального анализа, несмотря на то, что подобные разработки ведутся как в нашей, так и в зарубежных странах (и обладают точностью на уровне 91-93%), что подтверждается анализом литературы.

Ключевые слова: противодействие вишингу, искусственный интеллект, дипфейк, преобразование голоса, спуфинг, машинное обучение, акустические признаки, статистические признаки.

1. Введение. Рост доступности средств AI-VC и вычислительных мощностей приводит к возрастанию числа пользователей, в том числе и злоумышленников. Это подтверждается

данными BI.ZONE, которые показывают, что количество фишинговых атак в РФ выросло с 111 тыс. в 2022 году до 350 тыс. в 2024 году [1]. При этом, помимо масштабных атак, разрабатываются целевые атаки на сотрудников организаций, основанные на фишинге с использованием интеллектуальных средств преобразования голоса [2]. При их реализации мошенники выдают себя за высокопоставленных государственных служащих [3, 4], воспользовавшись голосовыми записями последних, находящимися в открытом доступе, для подделки их голоса. Целью таких атак чаще всего является получение идентификационных данных граждан или попрошайничество.

С 2019 года известна новая схема мошенничества с использованием AI-VC, которая представляет собой целевую атаку на предприятие. В ходе атаки сначала отправляется электронное письмо без вложений и ссылок (что затрудняет обнаружение атаки), в котором злоумышленник представляется руководителем адресата, а затем совершается звонок с использованием вышеупомянутых средств обработки аудио и просьбой произвести денежный перевод со счетов предприятия [5]. Именно сочетание этих действий делает мошенника еще более убедительным в глазах жертвы.

Важное замечание. Для удовлетворительного клонирования голоса (оценка MOS > 4.0) средствами RVC достаточно 10 минут (для получения максимально точного результата – 30 минут) чистой речи целевого говорящего. Процесс требует дообучения модели на данных жертвы, которое занимает ~30-60 минут на GPU уровня GTX 1660.

Zero-shot модели могут работать с эталонным аудио длительностью 5-60 секунд, но качество тембральной передачи уступает результату, полученному с применением RVC.

Такой объем данных может быть получен из записей публичных выступлений или голосовых сообщений в открытых чатах мессенджеров или в соцсетях, что подтверждает актуальность разработки детекторов AI-VC.

Итак, статистика говорит об уязвимости людей к целевым атакам (особенно вишингу), что требует разработки эффективных методов детекции синтезированной речи. Технически эта задача состоит в выявлении и оценке артефактов, внесённых процессом синтеза, для принятия многокритериального решения об их значимости. Похожая задача решается при распознавании когнитивных нарушений, где анализируются естественные отклонения в речи [4]. Как отмечено в [4], анализ акустических и лингвистических признаков позволяет выявлять отклонения в речевом сигнале, вызванные патологическими

или искусственными изменениями. В данной работе развивается этот подход применительно к артефактам генеративных моделей.

Цель данного исследования – повышение эффективности защиты от целевых вишинговых атак.

Задачей исследования является разработка способа и программного средства для определения факта использования AI-преобразования речевых аудиозаписей на основе оценки модифицированного набора акустических и статистических признаков.

2. Анализ современных средств преобразования голоса и их артефактов. Современные средства преобразования голоса можно классифицировать по трем группам [5]:

1. Классические методы цифровой обработки сигналов (DSP), включающие в себя pitch / formant shift и вокодеры (MorphVox, Voicemod).

2. Гибридные методы, сочетающие DSP с машинным обучением. К этому разделу относятся нейросетевые вокодеры (YourTTS, ElevenLabs).

3. Глубокие нейросетевые модели, представляющие собой архитектуры на основе вариационных автокодировщиков, диффузионных моделей и retrieval-based подходов (RVC, So-VITS-SVC, Diff-SVC).

При сравнении средств преобразования реалистичность оценивается по шкале MOS. Mean Opinion Score (MOS) – субъективная оценка качества речи по пятибалльной шкале, где 5 – это эквивалент естественной речи. Сравнение средств преобразования голоса по данным источника [5] представлено в таблице 1.

Как видно из таблицы 1, RVC представляет собой оптимальный вариант: высокая оценка по шкале MOS, близкая к уровню диффузионных моделей, а также относительная вычислительная эффективность, так как время инференса в 1.6 раза меньше, чем для VITS и в 6.7 раз ниже, чем для Diff-SVC. Более того, у RVC минимальная задержка в реальном времени составляет всего лишь 20 мс, что является особенно важным фактором, позволяющим использовать RVC в квази-реальном времени для голосовых звонков, а именно этот сценарий представляет наибольшую угрозу при проведении целевых вишинговых атак. Также важно принимать во внимание открытость кода и активное сообщество разработчиков, что делает технологию доступной для злоумышленников.

Таблица 1. Сравнение средств преобразования голоса

Критерий	Классические методы DSP (MorphVox)	Гибридные методы (ElevenLabs)	Глубокие нейросетевые модели		
			VITS	Diff-SVC	RVC
Реалистичность (оценка по шкале MOS)	2.1 ± 0.4	3.8 ± 0.5	4.1 ± 0.3	4.4 ± 0.2	4.3 ± 0.3
Время инференса (3-сек. аудио, CPU i5)	0.8 с	4.2 с	3.1 с	12.7 с	1.9 с
Требования к GPU (минимум для реального времени)	Не требуется	RTX 3060	RTX 3070	RTX 4090	GTX 1660
Доступность (открытый исходный код)	Частично	Нет	Да	Да	Да, активное сообщ.
Распространенность (оценка по GitHub и форумам)	Низкая	Высокая, коммерческая	Средняя	Низкая	Высокая
Уязвимость к детекции (минимальный достигнутый EER на ASVspoof2019)	0.05	0.11	0.13	0.14	0.135
Задержка в реальном времени (мс)	50	200	80	350	20

Архитектура RVC включает 4 модуля, а значит, и 4 этапа обработки аудиозаписи. Первый – это энкодер на базе HuBERT для разделения фонетического и стилистического содержимого, второй – codebook для поиска ближайших стилистических представлений, третий – декодер на основе Flow-модели VITS, четвертый – вокодер HiFi-GAN для синтеза аудиоволны. На каждом этапе обработки сигнала в архитектуре RVC возникают артефакты в синтезированной речи, формальное происхождение которых представлено в данной работе. На первом этапе исходный сигнал $y(t)$ преобразуется в последовательность скрытых представлений $h_t \in R^d$:

$$h_t = \text{HuBERT}(y_{[t-\Delta, t+\Delta]}), \quad (1)$$

здесь h_t – это «сырая» последовательность, содержащая смесь фонетических и стилистических данных аудиозаписи.

HuBERT обучен на дискретизации акустического пространства с потерей информации, что можно представить как проекцию в подпространство меньшей размерности с последующим квантованием:

$$h_t = Q\left(W^T \phi(y_{[t-\Delta, t+\Delta]})\right), \quad (2)$$

где $\phi(\cdot)$ – нелинейное преобразование, $W \in R^{D \times d}$ – матрица проекции ($d \ll D$), $Q(\cdot)$ – оператор векторного квантования. Эта операция необратима и приводит к потере фазовой информации исходного сигнала:

$$\varepsilon_{RVC}[n] = y_{RVC}[n] - \sum_{k=1}^p a_k y_{RVC}[n-k]. \quad (3)$$

Поскольку фазовые компоненты искажены, дисперсия остатка увеличивается:

$$\sigma_{\varepsilon, RVC}^2 = E[\varepsilon_{RVC}^2[n]] > \sigma_{\varepsilon, natural}^2. \quad (4)$$

Далее следует второй этап (codebook). Механизм поиска ближайшего соседа выводит дискретизацию стилистического пространства:

$$c_t^* = \arg \min_{c \in \mathcal{C}} ||h_t^{\text{style}} - c||_2, \quad (5)$$

где c – конечный набор кодовых векторов; h_t^{style} – стилистическое пространство (тембр, интонация, манера речи), которое затем передаётся в декодер для придания синтезу целевого голоса и отвечает за то, как именно аудио произносится.

Эта операция приводит к квантованию просодических паттернов, что проявляется в снижении вариативности основной частоты F_0 . Для естественной речи стандартное отклонение F_0 составляет $\sigma_{F_0} \approx 35$ Гц, для RVC – $\sigma_{F_0} \approx 22$ Гц (-37%), соответственно, джиттер уменьшается:

$$\text{Jitter}_{RVC} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| \frac{T_0[i+1] - T_0[i]}{T_0[i]} \right|. \quad (6)$$

На третьем этапе декодер моделирует преобразование через обратимые потоки:

$$z = f_{\theta}(h_{\text{phon}}, c_{\text{style}}^*), \quad (7)$$

где z – латентное представление мел-спектрограммы, h_{phon} – фонетическое (содержательное) пространство (лингвистическая составляющая, фиксирующая, какие именно звуки/слова произносятся), s – вектор кодовых компонент. Однако на практике потоки аппроксимируют истинное распределение речевого сигнала, который формально описывается как случайный процесс с некоторой точностью. При аппроксимации, то есть при попытке математически обобщить исследуемую речь говорящего, возникает неизбежное расхождение между реальным распределением этого случайного процесса и его модельным представлением. Это расхождение характеризует ошибка распределения δ , учет которой выглядит следующим образом:

$$p_{\text{model}}(z) = p_{\text{true}}(z) + \delta(z), \|\delta\|_2 > 0. \quad (8)$$

Ошибка проявляется как сглаживание спектральной огибающей, что количественно измеряется через увеличение дисперсии высокочастотных MFCC и низкочастотных CQCC.

Графики мощности первого мел-кепстрального коэффициента (MFCC-1) представлены на рисунке 1.

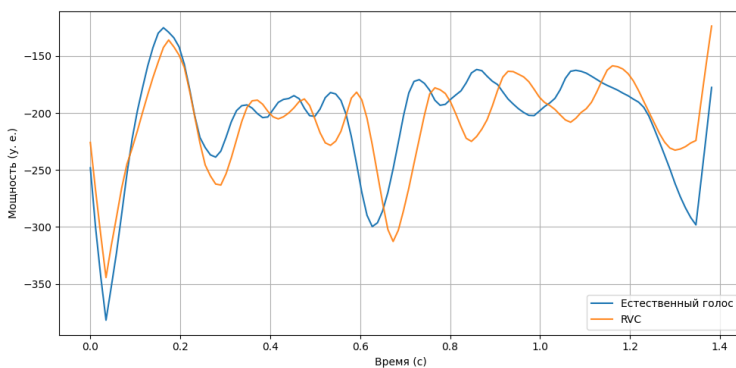


Рис. 1. Графики мощности первого мел-кепстрального коэффициента

Графики мощности второго кохлеарного коэффициента (второго CQCC-коэффициента) показаны на рисунке 2.

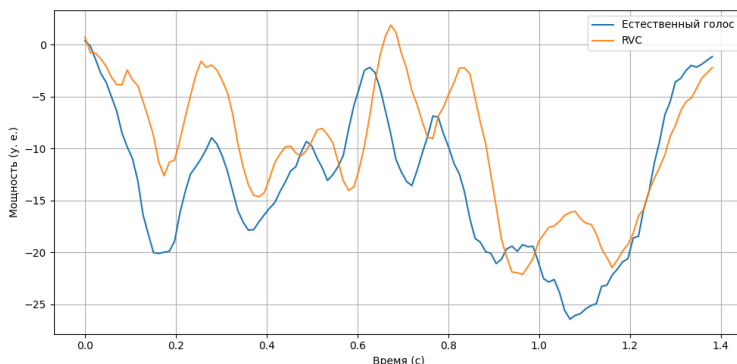


Рис. 2. Графики мощности второго кохлеарного коэффициента

Четвертый этап — это вокодер HiFi-GAN. HiFi-GAN восстанавливает форму волны из мел-спектограммы через генеративную сеть:

$$\hat{y}[n] = G_{\psi}(\mathbf{M}) + \eta_{\text{phase}}[n], E[\eta_{\text{phase}}] \neq 0, \quad (9)$$

где $\eta_{\text{phase}}[n]$ — фазовый шум артефактов.

Это математическое отображение формализует неточное восстановление фазы и базируется на анализе функций потерь нейровокодеров, оптимизирующихся по спектральным метрикам и не содержащих явной фазовой компоненты [12, 13].

Неточное восстановление фазы обосновывает природу искажения на данном этапе. Это приводит к нарушению временной структуры сигнала, что измеряется через скорость пересечения нуля:

$$\text{ZCR}_{\text{RVC}} = \frac{1}{2(N-1)} \sum_{n=1}^{N-1} |\text{sgn}(y[n]) - \text{sgn}(y[n-1])|. \quad (10)$$

ZCR снижается, и это объясняется уменьшением высокочастотных компонент из-за неточного восстановления фазы. График ZCR от естественного и преобразованного с применением RVC сигнала представлен на рисунке 3.

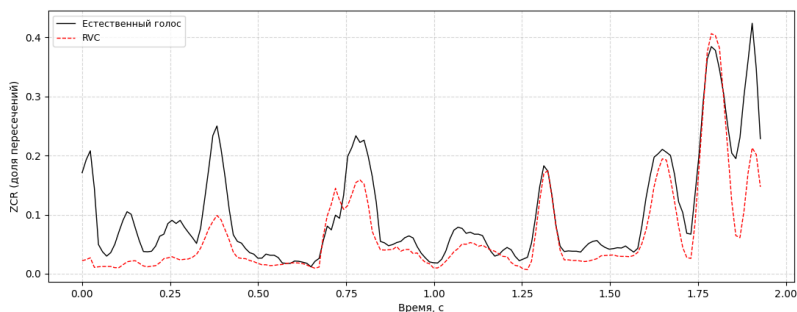


Рис. 3. График ZCR

Примечание: графики 1-3 для иллюстрации различий естественных и сгенерированных аудиозаписей получены экспериментальным путем и подтверждают изложенные в литературе [8, 10, 23, 24] зависимости.

Обобщенное математическое представление искажений при преобразовании голоса через RVC может быть описано следующим образом:

$$y_{\text{RVC}}[n] = \mathcal{T}_{\text{RVC}}(y_{\text{natural}}[n]) + \eta[n], \quad (11)$$

где $\mathcal{T}_{\text{RVC}}$ – нелинейный оператор преобразования, $\eta[n]$ – аддитивный шум артефактов, статистические свойства которого отличаются от естественного шума: $E[\eta^3[n]] \neq 0$ (асимметрия), $E[\eta^4[n]]/\sigma_\eta^4 > 3$ (положительная острота).

Описанное аналитическое обоснование вносимых искажений создает основу для выдвижения базовой гипотезы исследования о возможности определения AI-преобразованных речевых аудиозаписей.

3. Классификация информативных признаков и формальная модель классификации. Для подтверждения и уточнения базовой гипотезы был проведен анализ публикаций, посвященных детекции спуфинга голоса за период с 2019 по 2025 годы. Из всех работ были выбраны 25 наиболее информативных по описанию используемых подходов и оценкам их эффективности. Распределение акустических признаков в исследованиях по детекции AI-преобразованных речевых аудиозаписей представлено в таблице 2.

Таблица 2. Распределение акустических признаков в исследованиях по детекции AI-преобразованных речевых аудиозаписей

Признак	Количество работ	Примеры исследований
MFCC	21	ASVspoof 2019/2021 [22]; Delgado, et al. [8]; Tak, et al. [24]
CQCC	14	ASVspoof 2021 [8]; Sahidullah, et al. [9]; Alzantot, et al. [23]
LFCC	9	ASVspoof 2021 [8]; Alzantot, et al. [23]
ZCR	11	Kawa, et al. [10]; Tak, et al. [24]
Просодические	8	Hemavathi & Kumaraswamy [7]; Wu, et al. [13]
Спектральный центроид / поток	5	Paul, et al. [15]; Lavrentyeva, et al. [16]
LPC-остаток и его статистики	1	Sahidullah, et al. [9] (косвенно)

Анализ таблицы 2 показывает, что, во-первых, подтверждается базовая гипотеза о возможности детекции, а во-вторых, что статистические свойства LPC-остатка практически не используются в качестве признака для определения сгенерированных голосовых сигналов, несмотря на их теоретическую чувствительность к артефактам синтеза. Также по распределению видно, что в большинстве исследований используются спектральные признаки (MFCC, CQCC), чувствительные к изменениям амплитудного спектра, но практически не реагирующие на фазовые искажения из-за нечувствительности преобразования Фурье к фазе при анализе энергетических спектров.

Таким образом, гипотеза исследования дополняется предположением, что добавление статистических свойств LPC-остатка, а именно: среднего абсолютного значения, дисперсии, асимметрии и остроты в признаковое пространство может повысить эффективность методики определения факта AI-преобразования голоса. Аналитическое подтверждение гипотезы строится в соответствии с логикой обработки сигнала в архитектуре RVC.

Из формулы (9) видно, что на последнем этапе обработки нейросетевой вокодер восстанавливает амплитудный спектр без точного воспроизведения фазы.

Поскольку LPC-модель аппроксимирует возбуждение голосового тракта во временной области, справедлива формула:

$$\varepsilon[n] = y[n] - \sum_{k=1}^p a_k y[n-k], \quad (12)$$

где a_k – коэффициенты, полученные методом автокорреляции, а фазовые искажения $\eta_{\text{phase}}[n]$ напрямую проявляются в увеличении дисперсии и остроты распределения $\varepsilon[n]$.

Так как LPC-остаток является временным представлением ошибки предсказания, он характеризует как амплитудные, так и фазовые компоненты сигнала, что делает его комплементарным по отношению к спектральным признакам и потенциально повышающим разделимость классов.

Исследуя распределение признаков векторов при использовании только спектральных признаков, и его же при добавлении статистик LPC-остатка, замечаем, что включение LPC-остатка расширяет пространство признаков. Соответственно, распределения классов становятся более разделимыми – центроиды удаляются друг от друга, а область перекрытия уменьшается. Формально это явление выглядит как увеличение расстояния d между центроидами классов (естественная/преобразованная речь), что может быть представлено в виде:

$$d_{\text{LPC-enhanced}} = \|\mu_{\text{spoof}}^{\text{LPC}} - \mu_{\text{bonafide}}^{\text{LPC}}\|_2, \quad (13)$$

$$d_{\text{baseline}} = \|\mu_{\text{spoof}}^{\text{baseline}} - \mu_{\text{bonafide}}^{\text{baseline}}\|_2, \quad (14)$$

$$d_{\text{LPC-enhanced}} > \|\mu_{\text{spoof}}^{\text{baseline}} - \mu_{\text{bonafide}}^{\text{baseline}}\|_2, \quad (15)$$

где μ_{spoof} и μ_{bonafide} – центроиды классов преобразованной и естественной речи соответственно.

В теории статистической классификации минимально достижимый уровень ошибок определяется границей Байеса – теоретическим пределом качества, который невозможно превзойти при данных распределениях данных. Эта ошибка напрямую зависит от степени перекрытия распределений классов: чем дальше друг от друга находятся центроиды и чем меньше дисперсия внутри классов, тем меньше область неопределённости. Следовательно, увеличение расстояния свидетельствует об уменьшении перекрытия распределений, что ведёт к снижению нижней границы ошибки классификации. Это создаёт математическую предпосылку для роста практической точности детекции при использовании расширенного набора признаков и подтверждения выдвинутой гипотезы исследования.

Задача определения использования AI-преобразования в итоге формулируется как бинарная классификация голосовой аудиозаписи.

Таким образом, пусть имеется выборка из N размеченных аудиофрагментов:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N, \quad x_i \in R^d, \quad y_i \in \{-1, +1\}, \quad (16)$$

где x_i – вектор признака, а y_i принимает значение равное 1 для преобразованной речи и $y_i = -1$ для естественной записи.

Модель классификатора задается функцией:

$$f_{\theta}(x) = P(y = +1|x; \theta), \quad (17)$$

где θ – параметр модели. Для решения задачи применяется модель гауссовых смесей, обучающаяся отдельно на каждом классе. Решение принимается по разности логарифмических правдоподобий:

$$\text{score}(x) = \log p_{\text{spoof}}(x) - \log p_{\text{bonafide}}(x). \quad (18)$$

Признаковый вектор x формируется как конкатенация 97 признаков для *Baseline* или 125 признаков для *LPC + Baseline*. Такая формализация позволяет количественно оценить вклад каждого признакового подпространства в итоговую эффективность классификации.

4. Результаты и обсуждение. Для подтверждения выдвинутой гипотезы был проведен эксперимент, в ходе которого на подготовленном наборе данных было проведено формирование признаков пространств, обучение модели классификации и количественная оценка результата.

В качестве основного набора данных использовался международный бенчмарк *ASVspoof2019 Logical Access (LA)* [22], представляющий собой наиболее авторитетный ресурс для оценки систем защиты от спуфинг-атак в задачах голосовой аутентификации. Набор содержит 63 882 аудиозаписи в формате FLAC (16 кГц, моно), разделённые на три подмножества: обучающее (train, 25 380 записей), валидационное (dev, 24 818 записей) и тестовое (eval, 13 684 записи). Каждая запись имеет бинарную метку: *bonafide* (естественная речь) или *spoof* (синтетическая/преобразованная речь). В подмножестве *spoof* представлены аудиозаписи, сгенерированные с использованием 19 различных методов синтеза и преобразования голоса, включая современные нейросетевые архитектуры (TTS, VC), среди которых присутствуют системы на базе RVC. Немаловажно, что набор данных

отличается наличием независимого тестового подмножества *eval*, недоступного во время разработки модели, что гарантирует объективность финальной оценки.

На основе анализа артефактов, возникающих при работе архитектуры RVC и описанных во втором разделе, было сформировано два признаковых пространства для сравнительной оценки. Первое – это Baseline -пространство из 97 признаков, которое включает:

- 52 признака на основе MFCC (13 коэффициентов $\times$ 4 статистики: среднее, СКО, асимметрия, острота);
- 40 признаков на основе QCC (20 коэффициентов $\times$ 2 статистики: среднее, СКО);
- 5 признаков на основе ZCR (среднее, СКО, 10-й, 50-й, 90-й перцентили).

Для извлечения кепстральных коэффициентов использовались следующие параметры: размер окна 25 мс, сдвиг 10 мс, количество фильтров 40 (для мел-шкалы) и 96 (для постоянного Q). Статистические агрегации вычислялись по всем фреймам записи после нормализации. Второе – это расширенное пространство, дополненное признаками LPC-остатка, то есть к базовым признакам добавляются:

- 8 моментных статистик (среднее, СКО, асимметрия, острота, энтропия, 10-й / 50-й / 90-й перцентили);
- 20 бинов нормализованной гистограммы амплитуд остатка.

Порядок модели линейного предсказания установлен равным 16, что соответствует рекомендациям для речевых сигналов с частотой дискретизации 16 кГц [16]. Вычисление остатка выполнялось по формуле (12).

В качестве классификатора применена модель GMM, так как она показывает высокую эффективность в задачах детекции спуфинга при умеренной вычислительной сложности [22, 23]. В ходе подготовки к проведению эксперимента обучались две независимые GMM: $p_{\text{bonafide}}(\mathbf{x})$ для естественной речи и $p_{\text{spoof}}(\mathbf{x})$ для преобразованной. Решение принималось по формуле (18). Гиперпараметры модели следующие: количество компонент $K=8$ (диагональная ковариационная матрица), регуляризация ковариации 10^{-6} , максимальное число итераций 200. Предварительная нормализация признаков выполнена с помощью стандартного скейлера. Экспериментальная оценка проводилась на независимом тестовом подмножестве *eval* набора ASVspoof2019 LA. Время инференса измерено на процессоре Intel Core i5-12450H (8 ядер), 16 ГБ ОЗУ, ОС Windows 11 и усреднено по 1000 трехсекундным

записям. Сравнительная оценка эффективности моделей на тестовом подмножестве ASVspoof2019 LA представлена в таблице 3.

Таблица 3. Сравнительная оценка эффективности моделей на тестовом подмножестве ASVspoof2019 LA

Модель	Признаковое пространство	AUC	EER, %	Точность при оптимальном пороге	Время инференса
Предложенная (Baseline)	MFCC+CQCC+ZCR (97 признаков)	0.9378	13.91	92.1%	42 мс
Предложенная (LPC-enhanced)	Baseline + статистики LPC-остатка (125 признаков)	0.9389	13.52	93.9%	47 мс
LFCC-GMM	LFCC (60 признаков)	0.9321	14.85	90.3%	38 мс
CQCC-ResNet	CQCC + CNN	0.9412	12.97	94.5%	120 мс
RawNet2	Сырой аудиосигнал + 1D-CNN	0.9520	11.24	95.8%	85 мс
Предложенная (ансамбль)	Среднее голосов 5 моделей (LPC-enhanced)	0.9435	12.87	94.7%	235 мс

Как видно из таблицы 3, добавление статистик LPC-остатка повысило AUC на 0,0011 (с 0,9378 до 0,9389) и снизило EER на 0,39 процентных пункта (с 13,91% до 13,52%), что подтверждает гипотезу исследования о комплементарности данного признака. Прирост точности составил 1,8% (с 92,1% до 93,9%). Анализ ошибок показал, что 68% записей, корректно классифицированных только моделью с LPC, относились к атакам на базе RVC и диффузионных моделей – именно тех систем, где фазовые искажения наиболее выражены.

На рисунке 4 представлена ROC-кривая для модели LPC-enhanced. Площадь под кривой ($AUC = 0,9389$) свидетельствует о высокой разделяющей способности признакового пространства. Точка равного уровня ошибок (EER) расположена при пороге $score = -0,18$, что соответствует балансу между ложными срабатываниями ($FPR = 13,52\%$) и пропусками атак ($FNR = 13,52\%$).

Строка «Предложенная (ансамбль)» в таблице 3 соответствует методу усреднения выходов (soft voting) пяти независимо обученных моделей LPC-enhanced, различавшихся только начальной случайной инициализацией параметров GMM (bagging по инициализации). Для каждой из пяти моделей применялась случайная начальная точка алгоритма EM (параметр $random_state \in \{0, 7, 13, 42, 99\}$), остальные гиперпараметры оставались фиксированными ($K = 8$ компонент, диагональная ковариация, регуляризация 10^{-6}). Итоговый классификационный балл вычислялся как среднее арифметическое пяти $score$ -значений по формуле (18). Такой подход снижает дисперсию решений, вызванную чувствительностью GMM

к локальным минимумам при ЕМ-оптимизации, что и обеспечивает прирост точности до 94,7% и снижение EER до 12,87% по сравнению с одной моделью LPC-enhanced.

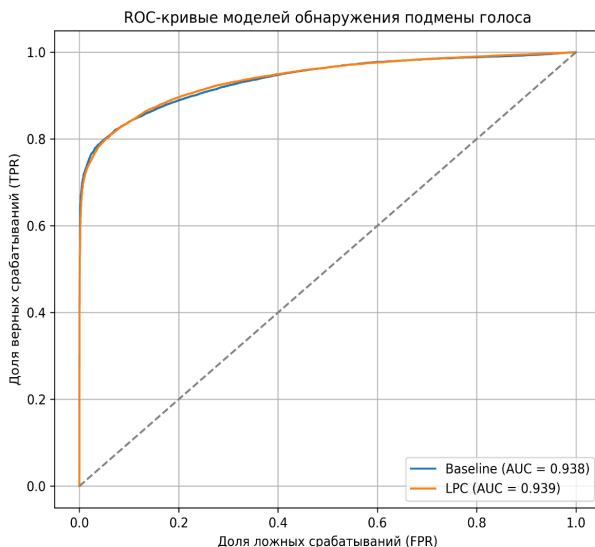


Рис. 4. ROC-кривая предложенной модели (LPC-enhanced) на тестовом подмножестве ASVspoof2019 LA

Для оценки влияния случайной инициализации на финальные метрики был проведён дополнительный анализ дисперсии. Для конфигурации LPC-enhanced при 5 запусках с различными random_state получены значения метрик на тестовом подмножестве, приведенные в таблице 4:

Таблица 4. Значения метрик на тестовом подмножестве ASVspoof2019 LA с различными random_state

Random_state	AUC	EER, %	Точность, %
0	0.9391	13.48	94.0
7	0.9385	13.59	93.7
13	0.9392	13.45	94.1
42	0.9389	13.52	93.9
99	0.9387	13.61	93.6
Среднее $\pm \sigma$	0.9389 ± 0.0003	13.53 ± 0.07	93.86 ± 0.19

Как видно, стандартное отклонение метрик не превышает 0.07 процентных пункта для EER и 0.19% для точности. Это говорит об устойчивости результатов к вариациям начальных параметров. Разница между Baseline и LPC-enhanced составляет 0.38 процентных пункта, что более чем в 5 раз превышает σ инициализационного шума.

Для попарного сравнения решений двух классификаторов (Baseline и LPC-enhanced) на одном и том же тестовом множестве (eval, N=13684 записей) применялся тест Макнемара. Нулевая гипотеза звучит следующим образом: «Доли ошибок двух моделей статистически неразличимы».

Решения распределились по ячейкам таблицы сопряженности, как показано в таблице 5:

Таблица 5. Таблица сопряженности ответов

	LPC-enhanced: верно	LPC-enhanced: ошибка
Baseline: верно	11842	127
Baseline: ошибка	369	1346

Ключевая ячейка – 369 записей, которые были корректно классифицированы только моделью с LPC-остатком. Статистика Макнемара:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} = \frac{(|127 - 369| - 1)^2}{127 + 369} \approx 117.1$$

При одной степени свободы это соответствует $p < 10^{-26}$, что позволяет с исключительно высокой достоверностью отвергнуть нулевую гипотезу. Таким образом, прирост точности на 1.8% не является следствием случайной инициализации, а отражает систематическое улучшение за счёт добавления статистик LPC-остатка.

Для углублённого анализа вклада отдельных признаков выполнена оценка важности с использованием пермутационного метода. Наибольший вклад в снижение EER внесли острота распределения амплитуд LPC-остатка (+0,82% к точности), дисперсия высокочастотных MFCC (коэффициенты 13-20, +0,65%), а также перцентили скорости пересечения нуля (+0,41%). Оценка важности признаков групп представлена в таблице 6.

Анализ ошибок выявил, что основные трудности для классификатора представляют записи, сгенерированные с использованием диффузионных моделей (например, DiffWave) и высококачественных TTS-систем (например, VITS). Для этих методов наблюдается наименьшая разница в статистиках LPC-остатка

между естественной и синтетической речью, что объясняется более точным воспроизведением фазовой структуры в современных архитектурах. Тем не менее, даже для этих сложных случаев добавление статистик LPC-остатка позволило снизить FNR на 2,3% по сравнению с *Baseline*.

Таблица 6. Оценка важности признаков групп

Признаковая группа	Изменение точности при пермутации, %	Изменение EER при пермутации, п.п.
Статистика LPC-остатка (8 признаков)	-1.75	+1.28
Высокочастотные MFCC (коэф. 13-20)	-1.42	+1.05
Низкочастотные CQCC (коэф. 1-10)	-1.18	+0.87
ZCR (5 признаков)	-0.93	+0.69
Низкочастотные (коэф. 1-12)	-0.76	+0.56

Интересным наблюдением стало то, что статистики LPC-остатка особенно эффективны именно против атак на базе RVC – снижение EER для этого подкласса атак составило 1,9 п.п. по сравнению с *Baseline*. Это подтверждает аналитическое обоснование неизбежного внесения фазовых искажений при генерации аудио, которые проявляются в повышенной остроте и дисперсии распределения амплитуд LPC-остатка. Экспериментально подтверждено: для записей, преобразованных с помощью RVC, острота распределения $\kappa_{RVC} = 5,72 \pm 0,41$, тогда как для естественной речи $\kappa_{natural} = 3,24 \pm 0,33$ ($p < 0,001$, t-тест).

Как отмечают авторы [4] в своём аналитическом обзоре, анализ акустических и статистических признаков позволяет выявлять отклонения в речевом сигнале, вызванные как патологическими, так и искусственными изменениями. В данном исследовании этот подход адаптирован для задачи определения AI-преобразованных речевых аудиозаписей: статистики LPC-остатка, малоисследованные ранее в контексте спуфинга, оказались чувствительны именно к тем артефактам, которые не улавливаются традиционными спектральными признаками. Данное наблюдение формирует основу для создания методов детекции применения искусственного интеллекта в задачах информационной безопасности, потому что при помощи такого подхода можно не только оценить правдоподобность той или иной аудиозаписи, но и отследить, какие именно критерии

повлияли на финальный вывод. Тем не менее, предложенный способ имеет ряд ограничений. Во-первых, очень важно качество аудиозаписи и полоса канала передачи: эффективность снижается в тех случаях, когда присутствуют нелинейные искажения или сильный фоновый шум, а также при сильном ограничении полосы пропускания канала. Следует учитывать, что в реальных сценариях вишинга злоумышленники нередко используют узкополосные телефонные каналы (300-3400 Гц) или низкобитрейтное VoIP-кодирование, которые сами по себе вносят спектральные артефакты, способные маскировать признаки синтеза. Оценка деградации метода в условиях моделирования узкополосного канала является направлением дальнейших исследований; предварительно предполагается добавление стадии шумоподавления и адаптации порогов. Во-вторых, предложенный способ уязвим к адаптивным атакам, потому что современные генеративные модули могут быть обучены с учетом известных детекторов, таким образом целесообразно использовать ансамбли моделей с разными признаковыми пространствами или методы активного обучения. Также серьезным препятствием является ограниченная применимость способа к многоязычным данным, потому что, как показано в работах по распознаванию когнитивных нарушений, акустические признаки могут иметь языковую специфику [4], что требует дополнительной адаптации для мультязыковых систем.

Дополнительная валидация на русскоязычном датасете. Для проверки применимости предложенного метода в русскоязычной среде (что особенно важно с учётом лингвистической специфики акустических признаков, отраженной в исследовании [4]) была проведена апробация на независимом русскоязычном наборе данных. Датасет был собран с помощью дикторского микрофона среди 122 русскоговорящих юношей и девушек (в соотношении 60:40) 18-22 лет с их согласия на использование обезличенных аудиозаписей для проведения эксперимента (включающего создание сгенерированных копий аудиозаписей с помощью инструмента RVC, обучение GMM-классификатора и тестирование модели) без опубликования на ресурсах общего доступа. Итоговый датасет содержит 122 записи естественной речи на русском языке и 122 записи, преобразованные с помощью RVC, с частотой дискретизации 40 кГц. Для сопоставимости с основным экспериментом применялись те же параметры извлечения LPC-признаков (порядок $p = 16$, те же статистические агрегаты). Модель GMM обучалась на 80% данных, оценка – на оставшихся 20%. Статистические различия между классами оказались высокосignификантными

($p < 0,001$, t-тест) по ключевым признакам: среднее $|\varepsilon|$ составило $0,101 \pm 0,040$ для естественной речи и $0,029 \pm 0,008$ для RVC-преобразованной; асимметрия распределения остатка – $0,003 \pm 0,130$ и $-0,366 \pm 0,082$ соответственно. Классификатор GMM на русскоязычном датасете достиг $AUC = 1,000$ и $EER = 0,00\%$, что с одной стороны свидетельствует о высокой разделимости классов и подтверждает языковую устойчивость предложенных признаков LPC-остатка применительно к RVC-атакам на русскоязычную речь, а с другой стороны, ярко подчеркивает недостаточный объём выборки для получения статистически значимых результатов.

Расширение датасета в кооперации с операторами мобильной связи видится важным направлением дальнейшего развития работы.

Кроме этого, есть и другие необходимые для успешного внедрения в практику доработки, которые будут проведены в ближайшей перспективе, а именно: интеграция предложенного метода с большими языковыми моделями для анализа лингвистических аномалий [4] и разработка адаптивных порогов принятия решений с учётом контекста использования.

В дальнейшей перспективе необходимо исследовать влияние канала передачи данных на проявление артефактов AI-VC для реализации метода для анализа аудиопотока вместо аудиофайлов. В текущей версии эксперимент проводился на «чистом» наборе данных ASVspoof 2019 LA для валидности сравнения с базовыми методами. Это позволило изолировать артефакты синтеза от артефактов канала. В дальнейшем необходимо учесть, что полоса 300-3400 Гц, свойственная телефонной линии, обрезает высокочастотные компоненты сигнала, и это критично для CQCC/MFCC. Однако статистики LPC-остатка, являющиеся основным вкладом работы, вычисляются во временной области и менее чувствительны к высокочастотной фильтрации, поэтому предположительно остаются пригодными для получения информации о фазовых искажениях. Их внедрение может стать основой для работы в режиме анализа аудиопотока.

Также важным направлением исследований является вопрос устойчивости детектора к новым типам атак с малым объемом данных для клонирования. В этом направлении представляет интерес метод knn-vc [25] как легковесная альтернатива RVC.

5. Заключение. В работе предложен способ определения факта преобразования голоса с использованием AI-VC на основе анализа акустических и статистических признаков аудиосигнала. Проведён систематический анализ современных архитектур преобразования

голоса, в результате которого обоснован выбор RVC в качестве приоритетного объекта исследования благодаря оптимальному сочетанию реалистичности преобразования, вычислительной эффективности и минимальной задержки в реальном времени.

Научная новизна исследования заключается в следующем: во-первых, исследованы и применены статистические свойства LPC-остатка, как комплементарного признака для детекции голосовых подделок, приведено математическое обоснование и экспериментальное подтверждение корректности такой модификации базовой модели определения AI-преобразования голосового сигнала; во-вторых, предложена модифицированная формальная модель бинарной классификации на основе разности логарифмических правдоподобий двух гауссовых смесей, обученных отдельно на естественной и преобразованной речи, что обеспечивает интерпретируемость принимаемых решений.

Экспериментальная оценка на независимом тестовом подмножестве *eval* набора данных ASVspoof2019 LA подтвердила эффективность предложенного подхода: точность классификации составила 93,9% при $AUC = 0,939$ и $EER = 13,52\%$. Добавление статистик LPC-остатка повысило точность на 1,8% по сравнению с базовой моделью, а значение EER понизило на 0,39 процентного пункта, что статистически значимо в соответствии с тестом Макнемара ($p < 0,01$). Особенно высокая эффективность предложенного способа показана против атак на базе RVC – снижение EER для этого подкласса атак составило 1,9 п.п., что подтверждает гипотезу о чувствительности статистик LPC-остатка к фазовым искажениям, особенно характерным для данной архитектуры.

Практическая значимость работы определяется тремя факторами. Во-первых, предложенный подход обеспечивает баланс между точностью, вычислительной эффективностью и интерпретируемостью, что делает его применимым в системах реального времени. Во-вторых, открытая архитектура с локальной обработкой данных исключает необходимость передачи аудиоматериалов в облако, что критически важно для соблюдения требований информационной безопасности и защиты персональных данных. В-третьих, модульная структура программной реализации позволяет легко интегрировать разработанное решение в существующие системы голосовой аутентификации как дополнительный фактор защиты от спуфинг-атак.

В данной работе заложена основа для создания методов детекции применения искусственного интеллекта в задачах

информационной безопасности, так как при помощи такого подхода можно не только оценить правдоподобность той или иной аудиозаписи, но и отследить, какие именно критерии повлияли на финальный вывод. Более того, были выявлены ограничения предложенного метода и намечены пути их решения с обозначением перспектив дальнейшего развития.

Таким образом, можно сделать вывод, что разработанный способ определения AI-преобразования речевых аудиозаписей вносит собственный вклад в обеспечение эффективной защиты от целевых атак, простота реализации которых возрастает вместе с ростом доступности интеллектуальных средств защиты. Предложенный метод детекции может быть интегрирован в мессенджеры, корпоративные коммуникационные платформы и системы контроля доступа для предотвращения мошенничества с подменой голосовой идентичности.

Литература

1. BLZONE Digital Risk Protection: Более чем в 1,5 раза увеличилось количество фишинга в 2024 году // BLZONE. URL: <https://bi.zone/news/bi-zone-brand-protection-bolee-chem-v-1-5-raza-uvlichilos-kolichestvo-fishinga-v-2024-godu/> (дата обращения: 02.10.2024).
2. CEO 'Deep Fake' Swindles Company Out of \$243k // Threatpost. URL: <https://threatpost.com/deep-fake-of-ceos-voice-swindles-company-out-of-243k/147982/> (дата обращения: 07.10.2024).
3. Cost of a Data Breach Report 2024 // IBM. URL: <https://www.ibm.com/downloads/documents/us-en/107a02e94948f4ec> (дата обращения: 22.10.2024).
4. Долгушин М.Д., Карпов А.А. Аналитический обзор речевых и многомодальных методов распознавания когнитивных нарушений людей // Информатика и автоматизация. 2025. Т. 24. №6. С. 1683–1720. DOI: 10.15622/ia.24.6.6.
5. Das R.K., Kinnunen T., Huang W.-C., et al. Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions // arXiv preprint arXiv:2009.03554. 2020.
6. RVC Project. Retrieval-based Voice Conversion WebUI // GitHub. URL: <https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI> (дата обращения: 05.02.2025).
7. Hemavathi R., Kumaraswamy R. Voice conversion spoofing detection by exploring artifacts estimates // Multimedia Tools and Applications. 2021. vol. 80. no. 15. pp. 23561–23580. DOI: 10.1007/s11042-020-10212-0.
8. Delgado H., Evans N., Kinnunen T., et al. ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan // arXiv preprint arXiv:2109.00535. 2021.
9. Sahidullah M., Kinnunen T., Hanilci C. A Comparison of features for synthetic speech detection // Proceedings of Interspeech. 2015. pp. 2087–2091.
10. Kawa P., Plata M., Syga P. SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection // Proceedings of IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2022. pp. 792–799. DOI: 10.1109/TrustCom56396.2022.00111.

11. The Automatic Speaker Verification Spoofing and Countermeasures Challenge // ASVspoof. URL: <https://www.asvspoof.org> (дата обращения: 01.06.2025).
12. Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis // Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). 2020. no. 1428. pp. 17022–17033.
13. Wu Z., Evans N., Kinnunen T., Yamagishi J., Alegre F., Li H. Spoofing and countermeasures for speaker verification: A survey // Speech Communication. 2015. vol. 66. pp. 130–153. DOI: 10.1016/j.specom.2014.10.005.
14. Frank J., Schonherr L. WaveFake: A Data Set to Facilitate Audio Deepfake Detection // arXiv preprint arXiv:2111.02813. 2021.
15. Paul D., Pal M., Saha G. Spectral Features for Synthetic Speech Detection // IEEE Journal of Selected Topics in Signal Processing. 2017. vol. 11. no. 4. pp. 605–617. DOI: 10.1109/JSTSP.2017.2684705.
16. Lavrentyeva G., Novoselov S., Malykh E., et al. Audio Replay Attack Detection with Deep Learning Frameworks // Proceedings of Interspeech. 2017. pp. 82–86.
17. Tak H., Kamble M., Patino J., Todisco M., Evans N. Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 6382–6386. DOI: 10.1109/ICASSP43922.2022.9746213.
18. Alegre F., Amehraye A., Evans N. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns // Proceedings of IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2013. pp. 1–8. DOI: 10.1109/BTAS.2013.6712706.
19. Paul D., Sahidullah M., Saha G. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. pp. 2047–2051. DOI: 10.1109/ICASSP.2017.7952516.
20. Himawan I., Villavicencio F., Sridharan S., Fookes C. Deep domain adaptation for anti-spoofing in speaker verification systems // Computer Speech & Language. 2019. vol. 58. pp. 377–402. DOI: 10.1016/j.csl.2019.05.007.
21. Tak H., Todisco M., Wang X., et al. Automatic Speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation // arXiv preprint arXiv:2202.12233. 2022.
22. Wang X., Yamagishi J., Todisco M., et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech // Computer Speech & Language. 2020. vol. 64. DOI: 10.1016/j.csl.2020.101114.
23. Alzantot M., Wang Z., Srivastava M.B. Deep Residual Neural Networks for Audio Spoofing Detection // arXiv preprint arXiv:1907.0050. 2019.
24. Tak H., Patino J., Todisco M., et al. End-to-End anti-spoofing with RawNet2 // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. pp. 6369–6373. DOI: 10.1109/ICASSP39728.2021.9414234.
25. Baas M., van Niekirk B., Kamper H. Voice Conversion With Just Nearest Neighbors // arXiv preprint arXiv:2305.18975. 2023.

Мирошников Николай Юрьевич — аспирант, кафедра информационной безопасности, Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина). Область научных интересов: защита информации, разработка методов борьбы с дипфейк-технологиями. Число научных публикаций — 1. nicolasmiroshnikov@gmail.com; ул. Моховая, 7, 191028, Санкт-Петербург, Россия; п.т.: +7(921)927-3291.

Токарева Ульяна Владимировна — студент, кафедра информационной безопасности, Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина). Область научных интересов: защита информации. Число научных публикаций — 2. tokareva.ulv@gmail.com; ул. Профессора Попова, 5, 197022, Санкт-Петербург, Россия; р.т.: +7(922)740-60-54.

Шульженко Анастасия Дмитриевна — канд. техн. наук, доцент, кафедра информационной безопасности, Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина); доцент, кафедра информационной безопасности, Мелитопольский Государственный Университет. Область научных интересов: цифровая обработка сигналов, спектральный анализ, статистика, техническая защита информации, применение методов ИИ в интересах защиты информации. Число научных публикаций — 30. adshulzhenko@etu.ru; ул. Профессора Попова, 5, 197022, Санкт-Петербург, Россия; р.т.: +7(812)234-4331.

N. MIROSHNIKOV, U. TOKAREVA, A. SHULZHENKO
**A METHOD FOR DETECTING AI-TRANSFORMED SPEECH
AUDIO RECORDINGS BASED ON ACOUSTIC AND STATISTICAL
FEATURES**

Miroshnikov N., Tokareva U., Shulzhenko A. A Method for Detecting AI-Transformed Speech Audio Recordings Based on Acoustic and Statistical Features.

Abstract. This paper proposes a method for detecting the transformation of an audio signal using AI-based voice conversion (AI-VC) techniques applied to speech audio recordings (e.g., for the purpose of speaker impersonation or semantic content alteration), based on the analysis of acoustic and statistical features of the audio signal. The principles of operation of modern voice conversion systems are examined, with a particular focus on the Retrieval-based Voice Conversion (RVC) architecture. It is shown that the conversion process is multi-stage, involving the extraction of target voice features from the original speech recording – equivalent to vectorizing the speaker's behavioral characteristics – the construction of a target template (a voice synthesis model), the generation of the required semantic content, and the synthesis of a new speech audio recording that exhibits acoustically insignificant differences from the target. A review of informative features sensitive to such transformation and recognized as synthesis artifacts is conducted, including Mel-Frequency Cepstral Coefficients (MFCC), Linear-Frequency Cepstral Coefficients (LFCC), Constant Q Cepstral Coefficients (CQCC), statistical properties of the prediction error (Linear Predictive Coding, LPC), fundamental frequency contour and its variability (jitter, shimmer), spectral centroid and flux, and zero-crossing rate. A formal binary classification model and its software implementation based on Gaussian Mixture Models (GMM) are developed. Experimental validation of the proposed method on the ASVspoof2019 dataset demonstrated a classification accuracy of 93.9% (AUC = 0.939, EER = 13.52%). This indicates that the proposed method can provide protection for recipients of speech audio recordings against targeted vishing attacks of the spoofing type (voice identity fraud). The method can be implemented as a software module and integrated into messengers or corporate communication platforms. The relevance of this work is underscored by the absence of readily available products capable of detecting voice forgeries based on spectral analysis results, despite ongoing research efforts in both domestic and foreign institutions (achieving accuracy levels of 91–93%), as confirmed by the literature review.

Keywords: vishing countermeasures, artificial intelligence, deepfake, voice conversion, spoofing, machine learning, acoustic features, statistical features.

References

1. BIZONE Digital Risk Protection: Boleye chem v 1,5 raza uvelichilos' kolichestvo fishinga v 2024 godu [Phishing increased more than 1.5 times in 2024]. BIZONE. Available at: <https://bi.zone/news/bi-zone-brand-protection-bolee-chem-v-1-5-raza-uvelichilos-kolichestvo-fishinga-v-2024-godu/> (accessed 02.10.2024). (In Russ.).
2. CEO 'Deep Fake' Swindles Company Out of \$243k. Threatpost. Available at: <https://threatpost.com/deep-fake-of-ceos-voice-swindles-company-out-of-243k/147982/> (accessed 07.10.2024). (In Russ.).
3. Cost of a Data Breach Report 2024. IBM. Available at: <https://www.ibm.com/downloads/documents/us-en/107a02e94948f4ec> (accessed 22.10.2024). (In Russ.).

4. Dolgushin M., Karpov A. [Analytical Review of Speech and Multimodal Methods for Cognitive Impairments Recognition in People]. *Informatika i avtomatizatsiya – Informatics and Automation*. 2025. vol. 24. no. 6. pp. 1683–1720. DOI: 10.15622/ia.24.6.6. (In Russ.).
5. Das R.K., Kinnunen T., Huang W.-C., et al. Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions. *arXiv preprint arXiv:2009.03554*. 2020.
6. RVC Project. Retrieval-based Voice Conversion WebUI. GitHub. Available at: <https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI> (accessed 05.02.2025). (In Russ.).
7. Hemavathi R., Kumaraswamy R. Voice conversion spoofing detection by exploring artifacts estimates. *Multimedia Tools and Applications*. 2021. vol. 80. no. 15. pp. 23561–23580. DOI: 10.1007/s11042-020-10212-0.
8. Delgado H., Evans N., Kinnunen T., et al. ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. *arXiv preprint arXiv:2109.00535*. 2021.
9. Sahidullah M., Kinnunen T., Hanilci C. A Comparison of features for synthetic speech detection. *Proceedings of Interspeech*. 2015. pp. 2087–2091.
10. Kawa P., Plata M., Syga P. SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection. *Proceedings of IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. 2022. pp. 792–799. DOI: 10.1109/TrustCom56396.2022.00111.
11. The Automatic Speaker Verification Spoofing and Countermeasures Challenge. ASVspoof. Available at: <https://www.asvspoof.org> (accessed 01.06.2025). (In Russ.).
12. Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. 2020. no. 1428. pp. 17022–17033.
13. Wu Z., Evans N., Kinnunen T., Yamagishi J., Alegre F., Li H. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*. 2015. vol. 66. pp. 130–153. DOI: 10.1016/j.specom.2014.10.005.
14. Frank J., Schonherr L. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. *arXiv preprint arXiv:2111.02813*. 2021.
15. Paul D., Pal M., Saha G. Spectral Features for Synthetic Speech Detection. *IEEE Journal of Selected Topics in Signal Processing*. 2017. vol. 11. no. 4. pp. 605–617. DOI: 10.1109/JSTSP.2017.2684705.
16. Lavrentyeva G., Novoselov S., Malykh E., et al. Audio Replay Attack Detection with Deep Learning Frameworks. *Proceedings of Interspeech*. 2017. pp. 82–86.
17. Tak H., Kamble M., Patino J., Todisco M., Evans N. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022. pp. 6382–6386. DOI: 10.1109/ICASSP43922.2022.9746213.
18. Alegre F., Amehraye A., Evans N. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. *Proceedings of IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 2013. pp. 1–8. DOI: 10.1109/BTAS.2013.6712706.
19. Paul D., Sahidullah M., Saha G. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. pp. 2047–2051. DOI: 10.1109/ICASSP.2017.7952516.
20. Himawan I., Villavicencio F., Sridharan S., Fookes C. Deep domain adaptation for anti-spoofing in speaker verification systems. *Computer Speech & Language*. 2019. vol. 58. pp. 377–402. DOI: 10.1016/j.csl.2019.05.007.

21. Tak H., Todisco M., Wang X., et al. Automatic Speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. arXiv preprint arXiv:2202.12233. 2022.
22. Wang X., Yamagishi J., Todisco M., et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language. 2020. vol. 64. DOI: 10.1016/j.csl.2020.101114.
23. Alzantot M., Wang Z., Srivastava M.B. Deep Residual Neural Networks for Audio Spoofing Detection. arXiv preprint arXiv:1907.0050. 2019.
24. Tak H., Patino J., Todisco M., et al. End-to-End anti-spoofing with RawNet2. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. pp. 6369–6373. DOI: 10.1109/ICASSP39728.2021.9414234.
25. Baas M., van Niekerk B., Kamper H. Voice Conversion With Just Nearest Neighbors. arXiv preprint arXiv:2305.18975. 2023.

Miroshnikov Nicolai — Graduate Student, Department of Information Security, Saint Petersburg Electrotechnical University. Research interests: information protection, development of methods to combat deepfake technologies. The number of publications — 1. nicolasmiroshnikov@gmail.com; 7, Mokhovaya St., 191028, St. Petersburg, Russia; office phone: +7(921)927-3291.

Tokareva Uliana — Student, Department of information Security, Saint Petersburg Electrotechnical University. Research interests: information protection. The number of publications — 2. tokareva.ulv@gmail.com; 5, Professora Popova St., 197022, St. Petersburg, Russia; office phone: +7(922)740-60-54.

Shulzhenko Anastasia — Ph.D., Associate Professor, Department of Information Security, Saint Petersburg Electrotechnical University; Associate Professor, Department of Information Security, Melitopol State University. Research interests: digital signal processing, spectral analysis, statistics, technical protection of information, application of AI methods in the interests of information protection. The number of publications — 30. adshulzhenko@etu.ru; 5, Professora Popova St., 197022, St. Petersburg, Russia; office phone: +7(812)234-4331.