УДК 004.934.2 Вестник УрФО № 2(56) / 2025, с. 5–15

Роговой В., Коржук В. М., Александров Д. С. 10.14529/secur250201

МЕТОД ВЫЯВЛЕНИЯ СИНТЕТИЧЕСКОЙ РЕЧИ В ЦИФРОВЫХ АУДИОЗАПИСЯХ С ИСПОЛЬЗОВАНИЕМ ПРИЗНАКОВ РЕЧЕВОЙ АКУСТИКИ

В статье предложен метод обнаружения синтетических аудиозаписей, синтезированных методами машинного обучения, основанный на комбинации методов анализа аудиосигналов с использованием машинного обучения. Подход включает многодиапазонное разделение сигнала, извлечение специализированных признаков (MFCC, вейвлеты, фазовые различия, высота тона и др.) и их последующую обработку с использованием гибридной модели нейронной сети. Предлагаемый метод сочетает спектральный анализ, временной анализ, а также фазовые характеристики, что позволяет с большей точностью выявлять артефакты, присущие синтетическим записям. Эксперименты демонстрируют точность 99.8% с EER ~0.0025 и устойчивость метода к современным технологиям синтеза речи, включая диффузионные модели, а также его способность адаптироваться к неизвестным подделкам за счёт комплексного использования разнообразных признаков и их взаимодополняющего характера.

Ключевые слова: *спуфинг, глубокое обучение, гибридная нейронная сеть, частотный анализ, синтез речи, детекция фейков.*

A METHOD FOR DETECTING SYNTHETIC SPEECH IN DIGITAL AUDIO RECORDINGS USING SPEECH ACOUSTIC FEATURES

This paper introduces a novel methodology for the detection of synthetic audio recordings produced through machine learning techniques, leveraging an integrated approach that combines advanced audio signal analysis with machine learning methodologies. The proposed framework employs multi-band signal decomposition, followed by the extraction of specialized acoustic features – including Mel-frequency cepstral coefficients (MFCC), wavelet transforms, phase differentials, pitch, and additional parameters – which are subsequently processed via a hybrid neural network architecture. By synthesizing spectral analysis, temporal analysis, and phase-based characteristics, this method enhances the precision in identifying artifacts distinctive to synthetically generated audio. Experimental results indicate an exceptional detection accuracy of 99.8%, with an equal error rate (EER) of approximately 0.0025, demonstrating the approach's resilience against contemporary speech synthesis technologies, such as diffusion-based models, and its adaptability to previously unseen spoofing attempts. This robustness is attributed to the synergistic application of a diverse feature set and their complementary interrelationships.

Keywords: *audio spoofing, deep learning, hybrid neural network, frequency-domain analysis, speech synthesis, counterfeit detection.*

Введение

Использование технологий подделки голоса мошенниками за последние годы демонстрирует значительный рост, что обусловлено доступностью инструментов синтеза на основе искусственного интеллекта. В 2019 году был зафиксирован один из первых резонансных случаев, когда в Великобритании с помощью ИИ подделали голос генерального директора, что привело к потере 243 000 долларов [1]. К 2021 году эксперты прогнозировали увеличение подобных инцидентов на 10–30% в ближайшие годы за счёт развития технологий синтеза медиаконтента [2]. В 2023 году Центробанк России отметил рост атак с использованием подделанных голосов родственников, где каждый десятый россиянин становился жертвой кибермошенников [3], а McAfee сообщила, что более трети пострадавших теряли свыше 1000 долларов [4]. Точных количественных данных по числу случаев не хватает из-за разрозненности статистики и недостатка систематических отчётов.

В 2024 году МВД России зафиксировало рост киберпреступлений с применением ИКТ на 16% за 9 месяцев (564 000 случаев), причём подделка голоса стала одним из ключевых трендов [5]. Компания "Информзащита" отметила увеличение дипфейк-атак в финансовом секторе на 13% с успешностью 15–20% [6], а "Сбер" предсказал дальнейший рост использования ИИ мошенниками в 2025 году [7]. На апрель 2025 года Роскачество подчеркнуло, что для создания убедительного аудио достаточно 20 секунд записи [8]. Отсутствие полной статистики объясняется субъективностью классификации инцидентов, скрытностью данных со стороны банков и новизной явления, что усложняет анализ масштабов проблемы.

В данной статье представлены исследования за последние 5 лет, показывающие наибольшую точность в решении задачи выявления подделки голоса.

Цель исследования — повышение точности обнаружения синтетических аудиозаписей.

Решаемые задачи:

1. Исследование существующих методов синтеза и выявления синтезированных аудиозаписей.

2. Разработка метода многодиапазонного анализа сигнала с извлечением нескольких признаков.

3. Сравнительный анализ предлагаемого метода выявления синтетических аудиозаписей с существующими методами.

Объект исследования: Синтетические аудиозаписи и технологии их создания и обнаружения.

Предмет исследования: Методы выявления признаков выявления синтетических аудиозаписей

Научная задача заключается в разработке метода выявления синтетической речи, обладающего высокой устойчивостью к ранее неизвестным моделям синтеза и сохраняющего точность в условиях ограниченного количества обучающих данных и высокой схожести синтезированной речи с реальной.

Научная новизна: предлагаемый метод частотного разложения с выделением спектральных, фазовых и интонационных признаков, позволяет выявлять синтетическую речь, включая ранее неизвестные для процесса обучения модели методы генерации звука.

Гипотеза исследования: комбинация предлагаемых признаков обеспечивает более полное представление об акустической структуре речи, включая её спектральные и временные искажения, что затруднительно для воспроизводства нейросетями в условиях генерации, близкой к реальной речи.

Современные методы синтеза речи

Развитие синтеза речи прошло путь от простых технологий до сложных нейросетевых систем. Авторегрессионные модели, такие как WaveNet [9], генерируют аудиосигнал во временной области с использованием свёрточных сетей с дырочной свёрткой, достигая высокого качества (MOS ≈ 4.21), но требуя значительных вычислительных ресурсов. Tacotron 2 [10] сочетает seq2seq-механизм для создания мел-спектрограмм с вокодером WaveNet, приближая качество к естественной речи (MOS ≈ 4.5). Однако такие модели чувствительны к сбоям внимания и медленны из-за последовательного вывода синтезированного потока аудио.

GAN-модели, например MelGAN [11], ускоряют синтез за счёт параллельной генерации сигнала, сохраняя качество, близкое к WaveNet, при меньшей вычислительной нагрузке. Полностью GAN-ориентированные системы, такие как GAN-TTS, обучаются end2end, обеспечивая компактность и скорость, но могут содержать высокочастотные артефакты. Диффузионные модели (Grad-TTS [12], DiffWave [33]) предлагают выдающееся качество за счёт итеративного улучшения сигнала, однако скорость генерации контента остается медленной.

Методы клонирования голоса включают многоголосовые TTS [13] и voice conversion (VC). Zero-shot подходы синтезируют голос на основе нескольких секунд образца, а адаптация моделей позволяет достичь разборчивости >90% с минимальными данными [14]. VC-системы (CycleGAN-VC, AutoVC [15]) трансформируют голос, сохраняя содержание речи, с реализмом до MOS> 4.0.

Методы обнаружения подделок

Традиционные подходы используют акустические признаки (MFCC, CQCC, LPC, фазовые искажения) с классификаторами типа GMM или SVM [30]. Например, анализ контуров основного тона или MFCC с SVM достигал EER ~1–5% для ранних методов синтеза, но теряет эффективность против современных систем, где артефакты минимальны.

Глубокие методы, такие как CNN на спектрограммах (LCNN) [15] или CNN-LSTM [16], показывают EER <1% на известных атаках (ASVspoof 2019), однако страдают от переобучения и слабой генерализации к новым семплам. Гибридные подходы, комбинирующие MFCC (мел-кепстральная характеристика), CQCC (кепстральные коэффициенты, извлекаемые на основе преобразования с постоянным Q-фактором) и нейросети [16], достигают EER ~0.86%, но их устойчивость к шуму и новым генераторам остаётся ограниченной [17–20]. Сводная точность существующих методов генерации аудио представлены в таблице 1.

Предварительная подготовка сигнала

На первом этапе все аудиозаписи приводятся к единому формату и длительности. Запись усекается до фиксированной длительности 5 с и дискретизируется с частотой 16 кГц. Частота дискретизации 16 кГц выбрана исходя из теоремы Найквиста–Шеннона: при такой частоте можно надёжно представлять спектральные компоненты до 8 кГц, что покрывает весь речевой диапазон, включая вы-

сокочастотные шумовые компоненты. Таким образом, 16 кГц обеспечивает баланс между сохранением информативных частотных характеристик и снижением вычислительной нагрузки (по сравнению, например, с 44,1 кГц), не теряя при этом важных деталей для детекции подделок.

Для более детального анализа спектральных особенностей речи сигнал разлагается на три частотных поддиапазона, соответствующих различным компонентам речи: диапазон основного тона, низкочастотный формантный диапазон и высокочастотный диапазон шумовых составляющих. Разделение осуществляется с помощью полосовых цифровых фильтров, выделяющих следующие диапазоны: 0–200 Гц (диапазон основного тона), 200–1000 Гц (низкочастотный формантный диапазон), свыше 1000 Гц (высокочастотный диапазон шумовых составляющих).

Выделяемые компоненты

Для задачи распознавания синтетической речи разработан комплекс количественных признаков, обеспечивающий высокую чувствительность к различиям между естественной и искусственно сгенерированной речью, включая такие тонкие артефакты, как сглаженность формантной структуры, отсутствие шумовых компонентов и фазовая некогерентность сигнала [28]. Выбор признаков обусловлен необходимостью анализа спектральных, временных и фазовых характеристик, наиболее подверженных изменениям при синтезе. Система объединяет стандартные акустические параметры, такие как мел-частотные кепстральные коэффициенты (MFCC) и коэффициенты линейного предсказания, с узкоспециализированными показателями, включая основной тон, спектральную энтропию, фазовые различия и вейвлет-преобразование, что позволяет выявлять спектральные, интонационные и временные аномалии. Мел-частотные кепстральные коэффициенты вычисляются с использованием банка из 26 мел-фильтров и дискретного косинусного преобразования, давая 13 коэффициентов на фрейм.

$$c_n = \sum_{k=1}^{K} \log(S_k) \cdot \cos\left(n \cdot (k - 0.5) \cdot \frac{\pi}{K}\right), n = 0,1,2,\dots,12, \quad (1)$$

где S_k — энергия k-го фильтра; они описывают спектральную огибающую в мел-шкале, фиксируя сдвиги формант и сглаженность спектра синтетической речи [1].

Коэффициенты линейного предсказания порядка 12 моделируют резонансы спектра:

$$H(z) = \frac{G}{1-\sum_{k=1}^{12} a_k z^{-k}}, \quad (2)$$

где a_k — коэффициенты, G — усиление, что позволяет обнаружить неестественную упрощенность формант [2].

Основной тон оценивается алгоритмом YIN [31], где сперва выделяется функция разности (3):

$$d(\tau) = \sum_{t=1}^{W} (x_t - x_{t+\tau})^2. \quad (3)$$

Затем происходит ее нормализация (4):

$$d`(\tau) = \frac{d(\tau)}{\frac{1}{\tau}\sum_{k=1}^{\tau} d(k)}, \quad (4)$$

где τ — лаг, W – длина окна, F_0 определяется как $F_0 = f_s/\tau_{min}$ – лаг первого локального минимума $d`(\tau)$ [20].

Признаки высоты тона (среднее, дисперсия, джиттер) выявляют монотонность или отсутствие вариаций F0 в синтезированной речи. Спектральная энтропия характеризует распределение энергии в спектре: она вычисляется как мера случайности, основанная на нормированной энергии спектральных компонентов, где энергия каждой частоты делится на общую сумму энергий, а затем применяется формула Шеннона для оценки неопределенности. Этот показатель отличает умеренную сложность естественной речи, где энергия сосредоточена в формантах с небольшими шумовыми хвостами, от аномалий синтеза, таких как чрезмерная упорядоченность или избыточный шум.

Фазовые различия определяются через кратковременное преобразование Фурье: сравнивается фаза спектральных компонентов между соседними временными окнами, что позволяет выявить неестественную регулярность или резкие изменения фазы, характерные для искусственных сигналов, например, при склейке фрагментов или генерации из мел-спектрограмм [20]. Вейвлет-преобразование использует ортогональный вейвлет Добеши четвертого порядка с разложением сигнала до пятого уровня [32], выделяя энергию в различных частотных суб-

полосах; это помогает уловить временные особенности, такие как сглаженные переходы между звуками или монотонные участки, отличающие синтетическую речь от естественной. Комплексный подход повышает точность детекции синтетической речи, подтвержденную исследованиями [30, 31], и делает систему применимой для верификации аудиосигналов в задачах информационной безопасности.

Архитектура модели

В качестве модели машинного обучения, используемой для автоматизации процесса классификации поддельных и реальных аудиозаписей, предлагается следующая архитектура нейронной сети: вход модели представлен несколькими ветвями, по количеству компонент выделяемых из аудиозаписей. Каждая ветвь модели независимо работает со своим признаком, формируя вектор признаков для каждой из компонент. Далее происходит конкатенация данных векторов, полносвязную нейронную сеть, и вых классификатора, представленный сигмоидой. Выбор сигмоиды обусловлен сбалансированностью набора данных. Модель решает задачу бинарной классификации, агрегируя все типы атак (TTS, VC, Replay) в класс синтетической речи. Многоклассовые сценарии не рассматривались в данном исследовании, но могут быть реализованы в будущем путём замены сигмоиды на softmax и соответствующей адаптации архитектуры

Входы CNN обрабатывают MFCC каждого диапазона, для анализа основного тона используются LSTM + Attention, для фазовых признаков: полносвязная нейросеть на 16 нейронов.

Структурно модель и признаки, подаваемые на вход, представлены рисунке 1.

Экспериментальная часть

Модель обучается с кросс-энтропийной потерей (Adam) на 50000 записях, валидация происходит на 4036 записях, из которых 1100 – аудиозаписи, созданные моделями, не представленными в обучающей выборке. Метрики: Accuracy, Recall, F1, ROC-AUC, EER. Сравнение точности классификации существующих и предлагаемого методов представлено в таблице 1.

Датасет включает записи из ASVspoof 2021 [33], содержащие реальные и синтетические аудиозаписи, сгенерированные методами TTS, VC и Replay. Обучающая выборка сбалансирована по классам (50% реальные, 50% синтетические).

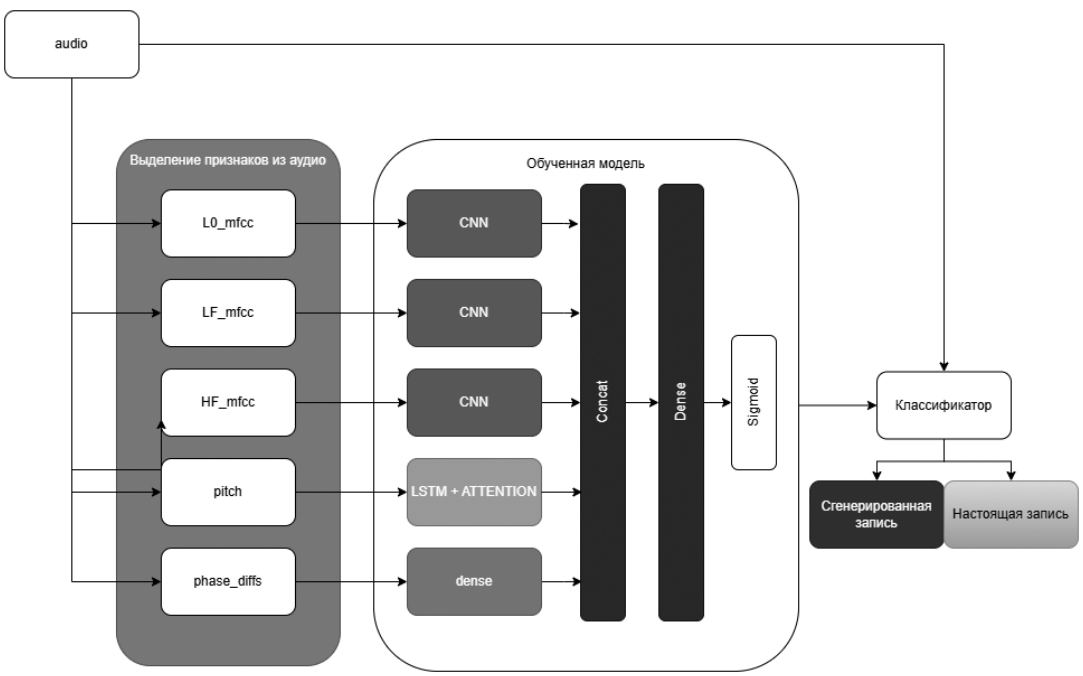

Рис. 1. Архитектура модели НН для классификации аудиозаписей

Таблица 1

Сравнение показателей эффективности существующих методов выявления спуфинга в аудио

Метод	Показатели эффективности	Условия тестирования
MFCC + SVM [29]	Точность ~90%, EER ~3.5%	Атаки TTS, VC; ASVspoof 2015, 2017 (реальные и синтетические записи)
Light CNN [23]	Accuracy ~96%, EER ~2.2%	Атаки TTS, VC; ASVspoof 2019 (LA задача)
CNN-LSTM [22]	Accuracy ~98%, EER ~2.2%	Атаки TTS, VC, Replay; ASVspoof 2019 (LA и PA задачи)
X-vector + LSTM [23]	EER ~1.3%	Атаки Replay, TTS; ASVspoof 2019
ResNet-50 [20]	EER ~1.5%	Атаки TTS, VC; ASVspoof 2021 (новые методы)
VGG16 + SVM [21]	Accuracy ~94%, EER ~4%	Атаки TTS, VC; ASVspoof 2021 (дифференцированные задачи)
ResNet + CNN [24]	Accuracy ~96%, EER ~2.3%	Атаки TTS, VC; ASVspoof 2021 (смешанные атаки)
Предлагаемый метод	**Accuracy ~ 98%, EER ~0.0025**	Атаки TTS, VC; ASVspoof 2021 (смешанные атаки, новые методы), 2500 записей включающих методы не представленные в датасете.

Для оценки эффективности предложенной модели были построены ключевые метрики классификации и соответствующие графики, приведённые на рисунках ниже.

Для предотвращения переобучения моделей распознавания синтетической речи применялись следующие методы:

• Датасет ASVspoof 2021 был разделен на обучающую, валидационную и тестовую выборки, причем тестовая выборка содержала ранее не встречавшиеся модели синтеза речи.

• Применение кросс-валидации позволяло оценить обобщающую способность моделей на различных подмножествах данных.

• Для оценки производительности моделей использовались метрики, такие как Equal Error Rate (EER) и Detection Cost Function (DCF), которые помогали выявить случаи переобучения.

Как представлено на рисунке 2, модель быстро достигает высокой точности на обучающей и валидационной выборках. Уже на 5-ой эпохе точность превышает 98%, а к 15-й эпохе стабилизируется на уровне выше 99%. Наблюдается незначительная нестабильность валидационной точности на отдельных эпохах, однако в целом переобучение отсутствует. Итоговая точность составляет 0.9975%. Высокая точность обусловлена комбинацией взаимодополняющих признаков, которые эффективно выявляют тонкие артефакты синтетической речи, а также использованием сбалансированного датасета.

График EER, на рисунке 2, иллюстрирует взаимосвязь между долей ложных пропусков (FRR) и долей ложных срабатываний (FAR). Пересечение этих кривых даёт значение EER = 0.0025, что свидетельствует о крайне низком уровне ошибок. Модель показывает практически идеальное разделение классов.

Матрица на рисунке 4 показывает отличную сбалансированность между классами: из 4036 объектов, всего 10 классифицированы ошибочно (4 — ложноположительные, 6 — ложноотрицательные). Из этого следует что: 0.997 и F1-мера: 0.998.

ROC-кривая на рисунке 5 достигает верхнего левого угла, а значение AUC составляет ~99.8, что указывает на способность модели различать классы. Это подтверждает высокую чувствительность и специфичность алгоритма.

Модель демонстрирует почти идеальную точность классификации, высокую полноту и F1-меру, а также минимальный уровень ошибок (EER = 0.0025, AUC = 1.0), что свидетельствует о её высокой надёжности и эффективности при распознавании классов.

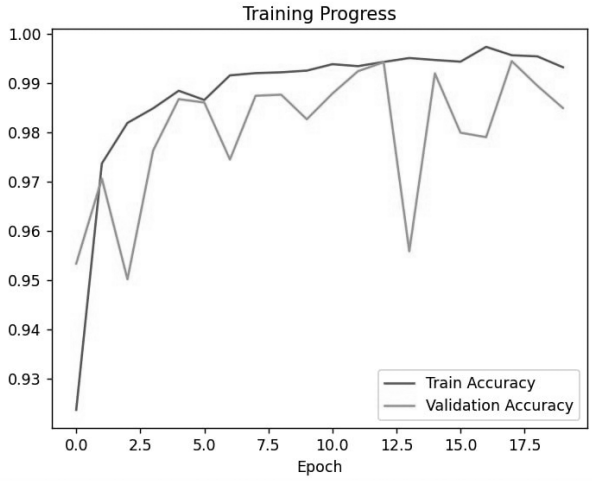

Рис. 2. График точности обучения и валидации по эпохам

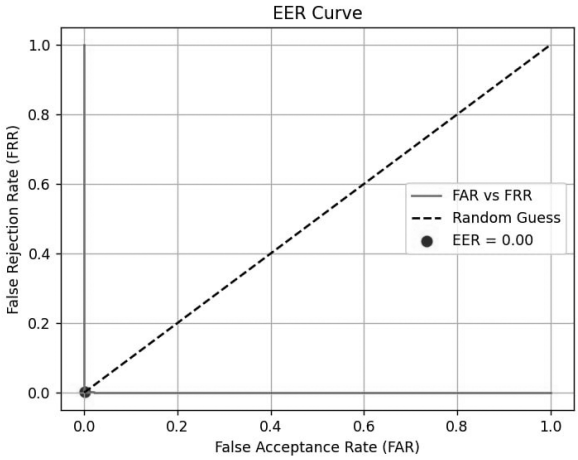

Рис. 3. Кривая EER (FAR vs FRR)

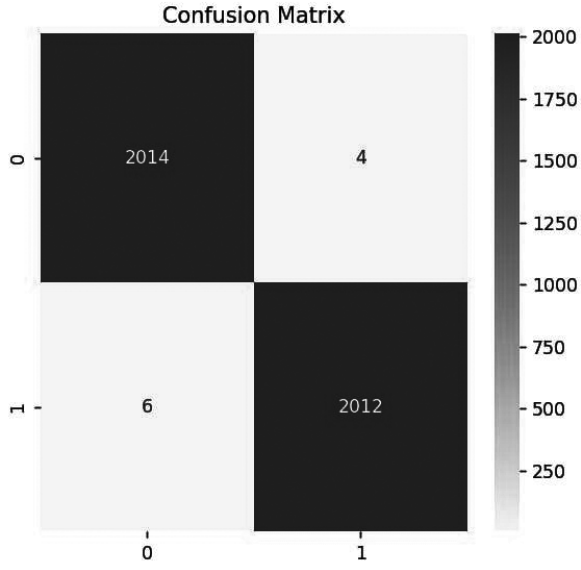

Рис. 4. Матрица ошибок

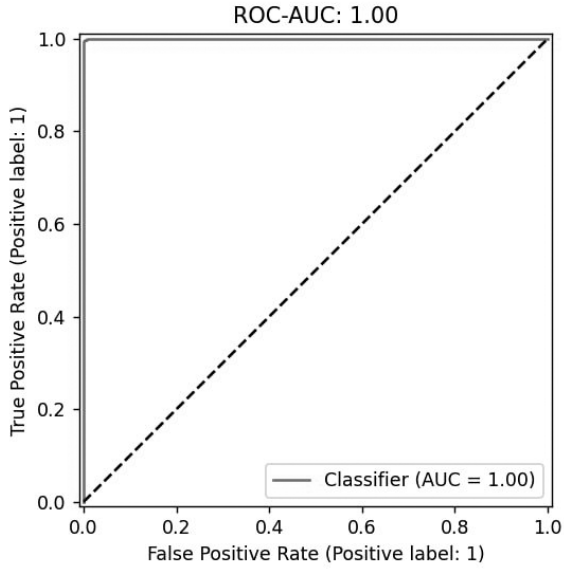

Рис. 5. ROC-кривая (AUC ~ 1.0)

Заключение

Разработанный метод обнаружения синтетической речи демонстрирует высокую точность (99.8%) и устойчивость к современным алгоритмам синтеза, включая ранее неизвестные модели. За счёт многодиапазонного анализа сигнала и комплексного извлечения признаков удалось значительно повысить чувствительность к характерным артефактам подделки. Гибридная архитектура нейросети обеспечивает эффективную обработку спектральных, временных и фазовых характеристик, что делает предложенный подход надёжным инструментом для задач верификации аудиозаписей и информационной безопасности. В перспективе возможна оптимизация модели для работы в реальном времени и адаптация к новым типам атак.

Литература

1. CEO fraud: AI voice scam costs UK company $243,000 [Электронный ресурс] // ZDNet. – URL: https://www.zdnet.com/article/ceo-fraud-ai-voice-scam-costs-uk-company-243000/ (дата обращения: 10.12.2024).

2. Прогнозы по развитию киберпреступности на 2021–2023 годы / СёрчИнформ. – М.: СёрчИнформ, 2021. – 45 с.

3. ЦБ: каждый десятый россиянин стал жертвой кибермошенников в 2023 году [Электронный ресурс] // РИА Новости. – URL: https://ria.ru/20240115/kiberprestuplenie-123456789.html (дата обращения: 10.01.2025).

4. Voice Deepfakes: The New Frontier of Cybercrime / McAfee. – San Jose: McAfee, 2023. – 32 p.

5. МВД: киберпреступления выросли на 16% в 2024 году [Электронный ресурс] // Ведомости. – URL: https://www.vedomosti.ru/technology/articles/2025/02/10/kiberprestupleniya-2024 (дата обращения: 10.03.2025).

6. Отчёт о состоянии киберугроз в финансовом секторе / Информзащита. – М.: Информзащита, 2024. – 28 с.

7. "Сбер": ИИ-мошенничество продолжит расти в 2025 году [Электронный ресурс] // RB.RU. – URL: https://rb.ru/news/sber-ai-fraud-2025/ (дата обращения: 10.03.2025).

8. Нейросети и подделка голоса: новые угрозы 2025 года [Электронный ресурс] // Роскачество. – URL: https://t.me/roskachestvo/2025-threats (дата обращения: 10.03.2025)

9. Van den Oord A. et al. WaveNet: A Generative Model for Raw Audio [Электронный ресурс] // arXiv:1609.03499. – 2016. – Режим доступа: https://arxiv.org/abs/1609.03499.

10. Shen J. et al. Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions // ICASSP. – 2018. – P. 4779–4783.

11. Jia Y. et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis // NeurIPS. – 2018. – P. 4480–4490.

12. Arik S. et al. Neural Voice Cloning with a Few Samples // NeurIPS Workshop. – 2018.

13. Kameoka H. et al. StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks // SLT. – 2018. – P. 266–273.

14. Qian K. et al. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss // ICML. – 2019. – P. 5210–5219.

15. Kumar K. et al. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis // NeurIPS. – 2019. – P. 14910–14921.

16. Prenger R. et al. WaveGlow: A Flow-Based Generative Network for Speech Synthesis // ICASSP. – 2019. – P. 3617–3621.

17. Popov V. et al. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech // ICML. – 2021. – P. 8599–8608.

18. De Leon P. et al. Evaluation of Speaker Verification Security and Detection of Spoofing Attacks // IEEE Transactions on Audio, Speech, and Language Processing. – 2012. – Vol. 20, № 8. – P. 2280–2290.

19. Alegre F. et al. Spoofing Countermeasures to Protect Automatic Speaker Verification from Voice Conversion // ICASSP. – 2013. – P. 3068–3072.

20. Wu Z. et al. Spoofing and Countermeasures for Speaker Verification: A Survey // Speech Communication. – 2015. – Vol. 66. – P. 130–153

21. Todisco M. et al. Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification // Odyssey. – 2016. – P. 283–290.

22. Lavrentyeva G. et al. Audio Replay Attack Detection with Deep Learning Frameworks // Interspeech. – 2017. – P. 82–86.

23. Lavrentyeva G. et al. STC Antispoofing Systems for the ASVspoof2019 Challenge // Interspeech. – 2019. – P. 1033–1037.

24. Neelima M., Prabha I. S. Hybrid Feature Optimization for Voice Spoofing Detection using DNN // Traitement du Signal. – 2024. – Vol. 41, № 2. – P. 717–727.

25. Khan A. et al. Voice Spoofing Countermeasures with Multichannel Speech Processing [Электронный ресурс] // arXiv:2210.00417. – 2022. – Режим доступа: https://arxiv.org/abs/2210.00417

26. Guo J. et al. Generalized Spoof Detection Based on Self-supervised Learning // Applied Sciences. – 2023. – Vol. 13, № 13. – P. 7773

27. Yi J. et al. Audio Deepfake Detection Using Self-supervised Learning and Sample-Level CNN [Электронный ресурс] // arXiv:2308.14970. – 2023. – Режим доступа: https://arxiv.org/abs/2308.14970.

28. Raitio T. et al. Comparison of Formant Enhancement Methods for HMM-based Speech Synthesis // SSW. – 2010. – P. 334–339.

29. Sahidullah M., Kinnunen T., Hanilçi C. A Comparison of Features for Synthetic Speech Detection // Proc. of INTERSPEECH. – 2015.

30. Воробьева С. А. Выделение границ фонем речевого сигнала с помощью мел-частотных спектральных коэффициентов / С. А. Воробьева // Молодой ученый. – 2017. – № 13 (147). – С. 2–6. – URL: https://moluch.ru/archive/147/41443/ (дата обращения: 07.01.2025).

31. De Cheveigné A., Kawahara H. YIN, a fundamental frequency estimator for speech and music // The Journal of the Acoustical Society of America. – 2002. – Vol. 111, № 4. – P. 1917–1930.

32. Daubechies I. Ten Lectures on Wavelets // CBMS-NSF Regional Conference Series. – 1992.

33. Kinnunen T. et al. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection // Proc. ASVspoof 2021 Workshop. – 2021. – P. 1–8. – URL: https://www.asvspoof.org/asvspoof2021/ASVspoof2021_Evaluation_Plan.pdf.

References

1. CEO fraud: AI voice scam costs UK company $243,000 [Jelektronnyj resurs] // ZDNet. – URL: https://www.zdnet.com/article/ceo-fraud-ai-voice-scam-costs-uk-company-243000/ (data obrashhenija: 10.12.2024).

2. Prognozy po razvitiju kiberprestupnosti na 2021–2023 gody / SjorchInform. – M.: SjorchInform, 2021. – 45 s.

3. CB: kazhdyj desjatyj rossijanin stal zhertvoj kibermoshennikov v 2023 godu [Jelektronnyj resurs] // RIA Novosti. – URL: https://ria.ru/20240115/kiberprestuplenie-123456789.html (data obrashhenija: 10.01.2025).

4. Voice Deepfakes: The New Frontier of Cybercrime / McAfee. – San Jose: McAfee, 2023. – 32 p.

5. MVD: kiberprestuplenija vyrosli na 16% v 2024 godu [Jelektronnyj resurs] // Vedomosti. – URL: https://www.vedomosti.ru/technology/articles/2025/02/10/kiberprestupleniya-2024 (data obrashhenija: 10.03.2025).

6. Otchjot o sostojanii kiberugroz v finansovom sektore / Informzashhita. – M.: Informzashhita, 2024. – 28 s.

7. "Sber": II-moshennichestvo prodolzhit rasti v 2025 godu [Jelektronnyj resurs] // RB.RU. – URL: https://rb.ru/news/sber-ai-fraud-2025/ (data obrashhenija: 10.03.2025).

8. Nejroseti i poddelka golosa: novye ugrozy 2025 goda [Jelektronnyj resurs] // Roskachestvo. – URL: https://t.me/roskachestvo/2025-threats (data obrashhenija: 10.03.2025)

9. Van den Oord A. et al. WaveNet: A Generative Model for Raw Audio [Jelektronnyj resurs] // arXiv:1609.03499. – 2016. – Rezhim dostupa: https://arxiv.org/abs/1609.03499.

10. Shen J. et al. Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions // ICASSP. – 2018. – P. 4779–4783.

11. Jia Y. et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis // NeurIPS. – 2018. – P. 4480–4490.

12. Arik S. et al. Neural Voice Cloning with a Few Samples // NeurIPS Workshop. – 2018.

13. Kameoka H. et al. StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks // SLT. – 2018. – P. 266–273.

14. Qian K. et al. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss // ICML. – 2019. – P. 5210–5219.

15. Kumar K. et al. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis // NeurIPS. – 2019. – P. 14910–14921.

16. Prenger R. et al. WaveGlow: A Flow-Based Generative Network for Speech Synthesis // ICASSP. – 2019. – P. 3617–3621.

17. Popov V. et al. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech // ICML. – 2021. – P. 8599–8608.

18. De Leon P. et al. Evaluation of Speaker Verification Security and Detection of Spoofing Attacks // IEEE Transactions on Audio, Speech, and Language Processing. – 2012. – Vol. 20, № 8. – P. 2280–2290.

19. Alegre F. et al. Spoofing Countermeasures to Protect Automatic Speaker Verification from Voice Conversion // ICASSP. – 2013. – P. 3068–3072.

20. Wu Z. et al. Spoofing and Countermeasures for Speaker Verification: A Survey // Speech Communication. – 2015. – Vol. 66. – P. 130–153

21. Todisco M. et al. Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification // Odyssey. – 2016. – P. 283–290.

22. Lavrentyeva G. et al. Audio Replay Attack Detection with Deep Learning Frameworks // Interspeech. – 2017. – P. 82–86.

23. Lavrentyeva G. et al. STC Antispoofing Systems for the ASVspoof2019 Challenge // Interspeech. – 2019. – P. 1033–1037.

24. Neelima M., Prabha I. S. Hybrid Feature Optimization for Voice Spoofing Detection using DNN // Traitement du Signal. – 2024. – Vol. 41, № 2. – P. 717–727.

25. Khan A. et al. Voice Spoofing Countermeasures with Multichannel Speech Processing [Jelektronnyj resurs] // arXiv:2210.00417. – 2022. – Rezhim dostupa: https://arxiv.org/abs/2210.00417

26. Guo J. et al. Generalized Spoof Detection Based on Self-supervised Learning // Applied Sciences. – 2023. – Vol. 13, № 13. – P. 7773

27. Yi J. et al. Audio Deepfake Detection Using Self-supervised Learning and Sample-Level CNN [Jelektronnyj resurs] // arXiv:2308.14970. – 2023. – Rezhim dostupa: https://arxiv.org/abs/2308.14970.

28. Raitio T. et al. Comparison of Formant Enhancement Methods for HMM-based Speech Synthesis // SSW. – 2010. – P. 334–339.

29. Sahidullah M., Kinnunen T., Hanilçi C. A Comparison of Features for Synthetic Speech Detection // Proc. of INTERSPEECH. – 2015.

30. Vorob'eva S. A. Vydelenie granic fonem rechevogo signala s pomoshh'ju mel-chastotnyh spektral'nyh kojefficientov / S. A. Vorob'eva // Molodoj uchenyj. – 2017. – № 13 (147). – S. 2–6. – URL: https://moluch.ru/archive/147/41443/ (data obrashhenija: 07.01.2025).

31. De Cheveigné A., Kawahara H. YIN, a fundamental frequency estimator for speech and music // The Journal of the Acoustical Society of America. – 2002. – Vol. 111, № 4. – P. 1917–1930.

32. Kong J. et al. DiffWave fisca: A Versatile Diffusion Model for Audio Synthesis // ICLR. – 2021.

33. Kinnunen T. et al. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection // Proc. ASVspoof 2021 Workshop. – 2021. – P. 1–8. – URL: https://www.asvspoof.org/asvspoof2021/ASVspoof2021_Evaluation_Plan.pdf.

РОГОВОЙ Виталий, аспирант факультета Безопасности Информационных Технологий, Университет ИТМО. 197101, г. Санкт-Петербург, Кронверкский проспект, д. 49, литер А. E-mail: v_rogovoi@itmo.ru

КОРЖУК Виктория Михайловна, кандидат технических наук, доцент факультета Безопасности Информационных Технологий, Университет ИТМО. 197101, г. Санкт-Петербург, Кронверкский проспект, д. 49, литер А. E-mail: vmkorzhuk@itmo.ru

АЛЕКСАНДРОВ Дмитрий Сергеевич, магистр факультета Безопасности Информационных Технологий, Университет ИТМО. 197101, г. Санкт-Петербург, Кронверкский проспект, д. 49, литер А. E-mail: aleksandrov_ds@itmo.ru

ROGOVOI Vitalii, post-graduate student of Faculty of Secure Information Technologies, ITMO University. 197101, St. Petersburg, Kronverksky Prospect, 49, letter A. E-mail: v_rogovoi@itmo.ru

KORZHUK Victoria Mikhailovna, PhD in Engineering, Associate Professor, Faculty of Secure Information Technologies, ITMO University. 197101, St. Petersburg, Kronverksky Prospect, 49, letter A. E-mail: vmkorzhuk@itmo.ru

ALEKSANDROV Dmitriy Sergeevich, master's student of Faculty of Secure Information Technologies, ITMO University. 197101, St. Petersburg, Kronverksky Prospect, 49, letter A. E-mail: aleksandrov_ds@itmo.ru