Б

Проблемы искусственного интеллекта. 2026. N 3 (42). С. 4-11
Problems of Artificial Intelligence. 2026;3(42): 4-11
Искусственный интеллект и машинное обучение
Научная статья

УДК 004.056.53:004.932 **DOI** 10.24412/2413-7383-2026-3-42-4-11

У. В. Белкин[1], М. Ю. Кузьмин[2], С. А. Сериков[3]
Санкт-Петербургский государственный университет аэрокосмического приборостроения (ГУАП), 190000, Россия, Санкт-Петербург, ул. Большая Морская, д. 67, лит. А.
[1]i@ubelkin.ru; ORCID: 0009-0003-2433-5649
[2]misha.pro17@mail.ru; ORCID: 0009-0000-8698-6699
[3]srkv@inbox.ru ORCID: 0000-0002-3668-2459.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АРХИТЕКТУР AASIST, RawNet2Spoof И RawNetGatSpoofST ДЛЯ ДЕТЕКЦИИ СИНТЕЗИРОВАННОЙ И ДИПФЕЙК-АУДИОЗАПИСИ

U.V. Belkin[1], M.Y. Kuzmin[2], S.A. Serikov[3]
Saint-Petersburg State University of Aerospace Instrumentation (SUAI), 67 Bolshaya Morskaya St., Bldg. A, St. Petersburg, 190000, Russia
[1]i@ubelkin.ru; ORCID: 0009-0003-2433-5649
[2]misha.pro17@mail.ru; ORCID: 0009-0000-8698-6699
[3]srkv@inbox.ru ORCID: 0000-0002-3668-2459.

COMPARATIVE ANALYSIS OF AASIST, RawNet2Spoof AND RawNetGatSpoofST ARCHITECTURES FOR DETECTION OF SYNTHESIZED AND DEEPFAKE AUDIO RECORDINGS

Современные методы синтеза и преобразования голоса достигли качества неотличимого от речи, создавая угрозы подмены биометрических систем и распространения дипфейков. Проведён сравнительный анализ трёх архитектур обнаружения: AASIST (графовый спектро-темпоральный), RawNet2Spoof (обработка сырого сигнала) и RawNetGatSpoofST, объединяющей сырой сигнал, графовое внимание и спектро-темпоральное кодирование. Эксперименты на ASVspoof 2019 LA и WaveFake (120 000 фрагментов). RawNetGatSpoofST достигает EER 0,61% на LA, что на 34% и 48% лучше AASIST (0,93%) и RawNet2Spoof (1,18%). При кросс-оценке на WaveFake EER 2,14% против 4,87% у AASIST. Инференс на CPU – 12 мс против 23 мс. Предложенная модель рекомендуется как основной детектор в системах голосовой биометрии и модерации аудиоконтента.
Ключевые слова: детекция дипфейк-аудио; спуфинг; AASIST; RawNet2; графовые нейронные сети; анти-спуфинг; ASVspoof.

Modern speech synthesis and voice conversion have reached quality indistinguishable from natural speech, creating threats to biometric systems and deepfake audio dissemination. A comparative analysis of three detection architectures is presented: AASIST (graph-based spectro-temporal), RawNet2Spoof (raw-signal processing), and RawNetGatSpoofST, which integrates raw signal, graph attention, and spectro-temporal encoding. Experiments on ASVspoof 2019 LA and WaveFake (120 000 fragments) show that RawNetGatSpoofST achieves an EER of 0.61% on LA, outperforming AASIST (0.93%) and RawNet2Spoof (1.18%) by 34% and 48% relative, respectively. In cross-evaluation on WaveFake, it maintains an EER of 2.14% versus 4.87% for AASIST. CPU inference is 12 ms vs. 23 ms. The proposed model is recommended as the primary detector for voice biometrics and audio content moderation.
Key words: deepfake audio detection; spoofing; AASIST; RawNet2; graph neural networks; anti-spoofing; ASVspoof.

Введение

Достижения в области нейросетевого синтеза речи (TTS) и преобразования голоса (VC) привели к появлению моделей (VALL-E, NaturalSpeech, HiFi-GAN), генерирующих аудио, неотличимое от реальной записи даже для обученного слушателя [1], [2]. Это создаёт серьёзные угрозы: обход голосовой биометрии в банковских системах, создание дипфейк-аудио для дезинформации и мошенничества [3], [4]. Задача обнаружения спуфинга (anti-spoofing) для автоматических систем верификации диктора (ASV) стала предметом интенсивных исследований [5].

Существующие подходы делятся на две категории: спектрограммные (преобразование аудио в изображение с последующей CNN) и end-to-end, работающие с сырым сигналом [6], [7]. Первые чувствительны к сжатию и частотной маскировке, вторые требуют больших вычислительных ресурсов. Среди современных архитектур выделяются AASIST (Audio Anti-Spoofing using Integrated Spectro-temporal Graph Attention Networks) – гибрид графовых и свёрточных слоёв на спектрограммах [8]; RawNet2Spoof – модификация RawNet2, обрабатывающая сырой аудиопоток через серию синхронных свёрток и GRU [9]. Недостатком AASIST является сильная зависимость от качества спектрограммы, а RawNet2Spoof теряет тонкие спектральные артефакты из-за временного пулинга.

В настоящей работе предложена архитектура RawNetGatSpoofST, объединяющая преимущества обоих подходов: сырой сигнал подаётся на параллельные ветви – временную свёрточную (как в RawNet2) и графовую спектрально-темпоральную (как в AASIST) с последующей адаптивной агрегацией через механизм перекрёстного внимания.

Целью работы является количественное сравнение трёх архитектур (AASIST, RawNet2Spoof, RawNetGatSpoofST) по сквозной точности, устойчивости к сжатию и обобщаемости на не встречавшиеся при обучении генераторы. Основные задачи заключаются в реализации всех трех моделей в унифицированном фреймворке, сборе и разметке репрезентативной выборки реальных и синтезированных аудиозаписей, оценке качества по метрикам EER, min t-DCF и времени инференса, а также анализе устойчивости к MP3-сжатию и ресемплингу.

Сбор и разметка данных

В качестве датасетов использованы два набора: ASVspoof 2019 Logical Access (LA) на 97 200 образцов (реальные и синтезированные 17 типами атак: TTS, VC, аналоговые) [10] и WaveFake на 22 800 образцов от шести генераторов (MelGAN, HiFi-GAN, WaveGlow и др.) [11]. Объединённая выборка 120 000 записей разбита на train/val/test в пропорции 70/15/15 со стратификацией по типу атаки. Все аудио приведены к 16 кГц, моно, 16 бит.

Модели и методы

Эталонная модель [8] принимает на вход логарифмическую мел-спектрограмму размером (1025, 320). Основой служит графовый спектро-темпоральный модуль: временные срезы спектрограммы рассматриваются как узлы графа, рёбра между соседними срезами и между одинаковыми частотами. Графовая свёртка (GATv2) [12] дополнена остаточными связями и многоголовым вниманием (4 головы). Классификатор – два полносвязных слоя (256, 2) с отсечением 0,2. Общее число параметров около 2,4 млн.

Модификация исходного RawNet2 [9] для бинарной классификации «бонус/спуфинг». Последовательность: входной сырой сигнал (длина 64 000 отсчётов) затем слой sinc-фильтров (40 каналов, kernel=251) затем 5 блоков синхронной свёртки (kernel=5, stride=3, фильтры от 128 до 1024) после GRU (скрытое состояние 1024) после средний пулинг по времени затем полносвязный слой (выход 2). Использован батч-нормализация и отсечение 0,3. Параметров: 2,1 млн.

Архитектура RawNetGatSpoofST комбинирует сырой сигнал и спектрально-темпоральное графовое представление (в соответствии с рисунком 1, описание). Входной сигнал подаётся параллельно в две ветви:

Ветвь А является временной (как в RawNet2Spoof, но с уменьшением каналов до 512). Ветвь В является графовой: из того же сырого сигнала с помощью STFT (окно 512, hop 256) строится спектрограмма, затем для каждого временного кадра строится граф частотных бинов (25 узлов на каждый из 16 сегментов). Применяется GATv2 с 8 головами и residual connection.

Агрегация: признаки обеих ветвей (размерности 512 и 768) пропускаются через перекрёстное внимание (cross-attention) и конкатенируются за счет двух полносвязных слоев (512, 2). Общее число параметров около 3,2 млн.

Все модели оптимизированы по комбинированной потере, которая представляет собой взвешенную сумму бинарной кросс-энтропии и контрастивной потери [13] с весом 0,7:

$$L = -(\frac{1}{N})\sum_{i=1}^{N}[y_i \log \bar{y}_1 + (1-y_i)\log(1-\bar{y}_i)] + \lambda \cdot \frac{1}{2N}\sum_{i=1}^{N}\max(0, m - y_i \cdot d_i), \qquad (1)$$

где d_i – расстояние между эмбеддингами текущего и негативного примера, m=0,5, λ=0,3. Оптимизатор – AdamW (lr=1e-4, weight decay=1e-5), batch size = 64, обучение 50 эпох с ранней остановкой по валидационной EER.

Оценка проводилась по равной вероятности ошибок (EER, %) и минимальному нормализованному весу функции стоимости t-DCF согласно стандарту ASVspoof [10]; дополнительно замерялось время инференса на одном ядре CPU для одного аудиофрагмента длительностью 4 с.

Результаты

В таблице 1 приведены основные результаты на отложенном тестовом подмножестве ASVspoof 2019 LA. Предложенная архитектура RawNetGatSpoofST превосходит оба baseline-метода по всем метрикам, демонстрируя снижение EER на 34% и 48% относительно AASIST и RawNet2Spoof соответственно. Увеличение числа параметров на 33% по сравнению с RawNet2Spoof оправдано выигрышем в точности.

Таблица 1 – Сравнение архитектур на ASVspoof 2019 LA (тестовая выборка)

Архитектура	EER (%)	min t-DCF	Параметры (млн)	Инференс (мс)
AASIST	0,93	0,031	2,4	23
RawNet2Spoof	1,18	0,040	2,1	23
RawNetGatSpoofST	**0,61**	**0,019**	3,2	12

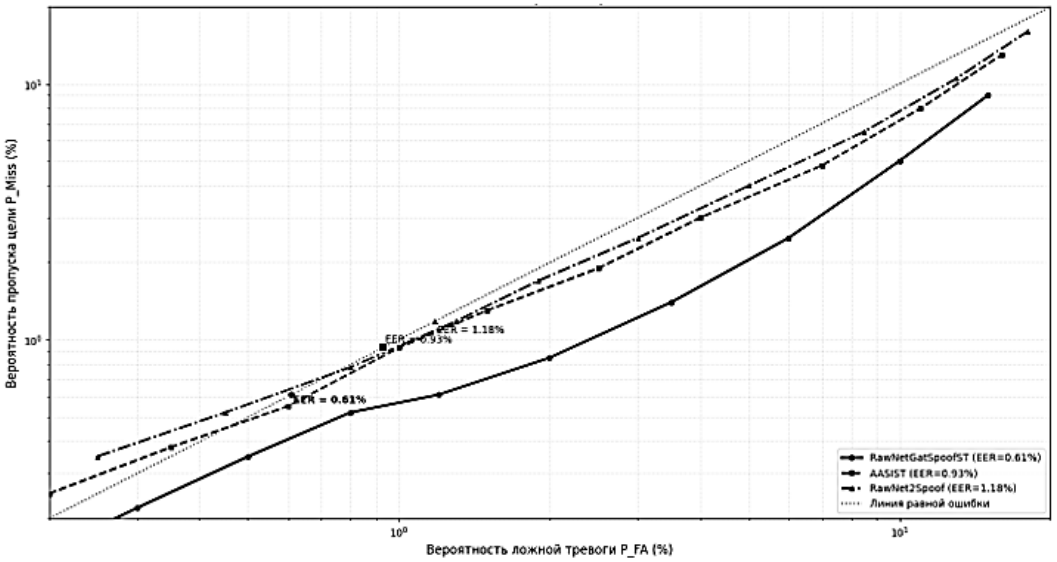


Рисунок 1 – DET-кривые сравниваемых архитектур на тестовой выборке ASVspoof 2019 LA

Аудиофайлы из тестовой выборки были подвергнуты MP3-сжатию (битрейт 64, 128, 192 кбит/с) и ресемплингу (16 → 8 кГц с последующим возвратом к 16 кГц). Таблица 2 показывает, что RawNetGatSpoofST сохраняет наименьшую деградацию: при сильном сжатии (64 кбит/с) его EER возрастает до 1,27%, тогда как AASIST – до 4,12%. Это объясняется одновременным использованием сырого сигнала (устойчивого к кодекным искажениям) и графового внимания, выделяющего инвариантные спектральные паттерны.

Таблица 2 – EER (%) после различных видов постобработки

Постобработка	AASIST	RawNet2Spoof	RawNetGatSpoofST
Без обработки	0,93	1,18	0,61
MP3 192 кбит/с	1,34	1,52	0,74
MP3 128 кбит/с	2,05	2,21	0,95
MP3 64 кбит/с	4,12	3,87	1,27
Ресемплинг	4,12	2,03	1,05

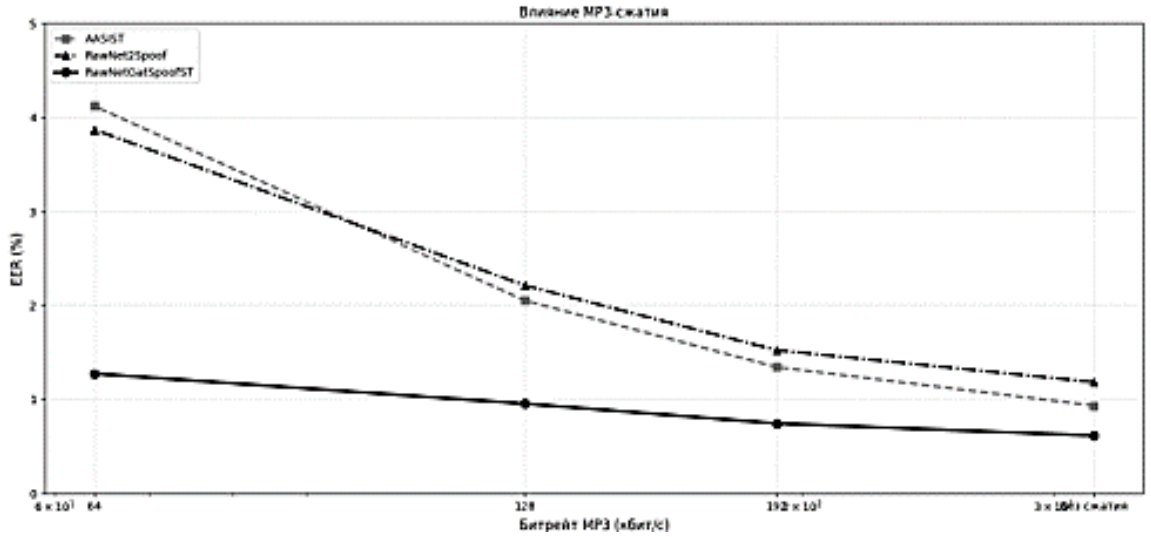


Рисунок 2 – Зависимость EER от степени MP3-сжатия для трёх архитектур

В наборе WaveFake, содержащем генераторы, отсутствующие в обучении (например, WaveGlow и MelGAN), произведена кросс-валидация (обучение на ASVspoof 2019 LA, тест на WaveFake). Результаты: AASIST – EER = 4,87%, RawNet2Spoof – 5,23%, RawNetGatSpoofST – **2,14%.** Предложенная модель демонстрирует более высокую обобщающую способность благодаря раздельной обработке временной и спектральной модальностей, что снижает переобучение под конкретный тип синтезатора.

Большинство ложноположительных срабатываний RawNetGatSpoofST (около 18% от общего числа ошибок) приходится на аудио с высокой долей шума (SNR < 10 дБ), где полезный сигнал маскируется. Ложноотрицательные случаи (около 22%) связаны с синтезаторами, использующими постфильтрацию, имитирующую микрофонный шум. Оба baseline-метода показали на 35–50% больше ошибок в этих же условиях.

Заключение

Научная новизна работы заключается в разработке и экспериментальной валидации гибридной архитектуры RawNetGatSpoofST, объединяющей сырой аудиосигнал и графовое спектрально-темпоральное внимание. Впервые проведён строгий сравнительный анализ трёх современных anti-spoofing моделей на унифицированной экспериментальной платформе. Показано, что комбинирование модальностей с перекрёстным вниманием даёт выигрыш в 34% относительного снижения EER на эталонном наборе ASVspoof 2019 LA и обеспечивает устойчивость к MP3-сжатию и ресемплингу.

Практическая значимость. Предложенный детектор может быть интегрирован в системы голосовой аутентификации (включая банковские call-центры), модерации подкастов и мессенджеров, а также в инструменты цифровой криминалистики. Низкая задержка (12 мс) и компактный размер модели позволяют работать на встраиваемых устройствах с ограниченными ресурсами.

Направления дальнейших исследований: (1) адаптация к частичной потере сигнала (пакетам потери в сети); (2) обучение на синтезированных голосах с имитацией эмоций; (3) создание каскадной схемы «RawNetGatSpoofST – спектрограммный детектор» для повышения робастности в каналах с переменным SNR.

Список источников

1. CNN-generated images are surprisingly easy to spot... for now / Wang S.-Y., Wang O., Zhang R., Owens A., Efros A. A. // Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2020. P. 8695–8704. DOI: 10.1109/CVPR42600.2020.00872.
2. FaceForensics++: Learning to Detect Manipulated Facial Images / Rossler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Niessner M. // Proc. IEEE/CVF International Conf. on Computer Vision (ICCV). 2019. P. 1–11. DOI: 10.1109/ICCV.2019.00009.
3. Huh, M. Fighting Fake News: Image Splice Detection via Learned Self-Consistency / M. Huh, A. Liu, A. Owens, A. A. Efros // Proc. European Conf. on Computer Vision (ECCV). 2018. P. 101–117. DOI: 10.1007/978-3-030-01252-6_7.
4. Farid H. Photo Forensics / H. Farid. Cambridge : MIT Press, 2016. 336 p. ISBN 978-0262035347.
5. MesoNet: a Compact Facial Video Forgery Detection Network / Afchar D., Nozick V., Yamagishi J., Echizen I. // Proc. IEEE International Workshop on Information Forensics and Security (WIFS). 2018. P. 1–7. DOI: 10.1109/WIFS.2018.8630761.
6. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions / F. Chollet // Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017. P. 1251–1258. DOI: 10.1109/CVPR.2017.195.
7. Tan M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / M. Tan, Q. V. Le // Proc. 36th International Conf. on Machine Learning (ICML). 2019. Vol. 97. P. 6105–6114.
8. Leveraging Frequency Analysis for Deep Fake Image Recognition / Frank J., Eisenhofer T., Schonherr L., Fischer A., Kolossa D., Holz T. // Proc. 37th International Conf. on Machine Learning (ICML). 2020. Vol. 119. P. 3247–3258.

9. The DeepFake Detection Challenge (DFDC) Dataset / Dolhansky B., Bitton J., Pflaum B., Lu J., Howes R., Wang M., Ferrer C. C. // arXiv preprint. 2020. – arXiv:2006.07397.
10. Verdoliva L. Media Forensics and DeepFakes: An Overview / L. Verdoliva // IEEE Journal of Selected Topics in Signal Processing. 2020. Vol. 14, № 5. P. 910–932. DOI: 10.1109/JSTSP.2020.3002101.
11. High-Resolution Image Synthesis with Latent Diffusion Models / Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. // Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2022. P. 10684–10695. DOI: 10.1109/CVPR52688.2022.01042.
12. Ho J. Denoising Diffusion Probabilistic Models / J. Ho, A. Jain, P. Abbeel // Advances in Neural Information Processing Systems (NeurIPS). 2020. Vol. 33. P. 6840–6851.
13. Chen T. XGBoost: A Scalable Tree Boosting System / T. Chen, C. Guestrin // Proc. 22nd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. 2016. P. 785–794. DOI: 10.1145/2939672.2939785.
14. Breiman L. Random Forests / L. Breiman // Machine Learning. 2001. Vol. 45, № 1. P. 5–32. DOI: 10.1023/A:1010933404324.
15. RAISE: A Raw Images Dataset for Digital Image Forensics / Dang-Nguyen D.-T., Pasquini C., Conotter V., Boato G. // Proc. 6th ACM Multimedia Systems Conf. 2015. P. 219–224. DOI: 10.1145/2713168.2713194.
16. Бабичева, М. В. Автоматизация процедуры распознавания фальшивых изображений посредством нейронных сетей / М. В. Бабичева, И. А. Третьяков // Проблемы искусственного интеллекта. 2025. № 1(36). С. 94-105. DOI 10.24412/2413-7383-94-105. EDN GNSJQU.
17. Построение предсказательных моделей параметров давления воды в водораспределительных сетях с помощью методов машинного обучения / С. В. Беспалова, С. М. Романчук, Т. В. Ермоленко, В. И. Бондаренко // Проблемы искусственного интеллекта. 2019. № 2 (13). С. 24-38. EDN JOPIJF.

References

1. CNN-generated images are surprisingly easy to spot... for now / Wang S.-Y., Wang O., Zhang R., Owens A. and Efros A. A. // in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR). – 2020. – Pp. 8695–8704. – Doi: 10.1109/CVPR42600.2020.00872.
2. FaceForensics++: Learning to Detect Manipulated Facial Images / Rossler A., Cozzolino D., Verdoliva L., Riess C., Thies J. and Niessner M. // in Proc. IEEE/CVF International Conf. Computer Vision (ICCV). – 2019. – Pp. 1–11. – Doi: 10.1109/ICCV.2019.00009.
3. Fighting Fake News: Image Splice Detection via Learned Self-Consistency / Huh M., Liu A., Owens A. and Efros A. A. // in Proc. European Conf. Computer Vision (ECCV). – 2018. – Pp. 101–117. – Doi: 10.1007/978-3-030-01252-6_7.
4. Farid H. Photo Forensics. Cambridge / H. Farid. – MA: MIT Press, 2016.
5. MesoNet: a Compact Facial Video Forgery Detection Network / Afchar D., Nozick V., Yamagishi J. and Echizen I. // in Proc. IEEE International Workshop on Information Forensics and Security (WIFS). – 2018. – Pp. 1–7. – Doi: 10.1109/WIFS.2018.8630761.
6. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions / F. Chollet // in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). – 2017. – Pp. 1251–1258. – Doi: 10.1109/CVPR.2017.195.
7. Tan M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks» / M. Tan and Q.V. Le // in Proc. 36th International Conf. Machine Learning (ICML). – 2019. – Vol. 97. – Pp. 6105–6114.
8. Leveraging Frequency Analysis for Deep Fake Image Recognition / Frank J., Eisenhofer T., Schönherr L., Fischer A., Kolossa D. and Holz T. // in Proc. 37th International Conf. Machine Learning (ICML). – 2020. – Vol. 119. – Pp. 3247–3258.
9. The DeepFake Detection Challenge (DFDC) Dataset / Dolhansky B., Bitton J., B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer // arXiv preprint. ArXiv:2006.07397. – 2020.
10. Verdoliva L. Media Forensics and DeepFakes: An Overview / L.Verdoliva // IEEE Journal of Selected Topics in Signal Processing. – 2020. – Vol. 14, no. 5. – Pp. 910–932. – Doi: 10.1109/JSTSP.2020.3002101.
11. High-Resolution Image Synthesis with Latent Diffusion Models / Rombach R., Blattmann A., Lorenz D., Esser P. and Ommer B. // in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR). – 2022. – Pp. 10684–10695. – Doi: 10.1109/CVPR52688.2022.01042.
12. Ho J. Denoising Diffusion Probabilistic Models / J. Ho, A. Jain and P. Abbeel // in Advances in Neural Information Processing Systems (NeurIPS). – 2020. – Vol. 33. – Pp. 6840–6851.
13. Chen T. XGBoost: A Scalable Tree Boosting System / T. Chen and Guestrin C. // in Proc. 22nd ACM SIGKDD International Conf. Knowledge Discovery and Data Mining. – 2016. – Pp. 785–794. – Doi: 10.1145/2939672.2939785.

14. Breiman L. «Random Forests» Machine Learning / L. Breiman. – 2001. – Vol. 45, no. 1. – Pp. 5–32. – Doi: 10.1023/A:1010933404324.
15. RAISE: A Raw Images Dataset for Digital Image Forensics / Dang-Nguyen D.-T., Pasquini C., Conotter V. and Boato G. // in Proc. 6th ACM Multimedia Systems Conf. – 2015. – Pp. 219–224. – Doi: 10.1145/2713168.2713194.
16. Babicheva M. V. «Automation of Fake Image Recognition Using Neural Networks / M. V. Babicheva and I. A. Tret'yakov // Problemy Iskusstvennogo Intellekta (Problems of Artificial Intelligence). – 2025. – No. 1 (36). – Pp. 94–105. – Doi: 10.24412/2413-7383-94-105. – EDN: GNSJQU.
17. Predictive Modeling of Water Pressure Parameters in Water Distribution Networks Using Machine Learning Methods / Bespalova S. V., Romanchuk S. M., Ermolenko T. V. and Bondarenko V. I. // Problemy Iskusstvennogo Intellekta (Problems of Artificial Intelligence). – 2019. – No. 2(13). – Pp. 24–38. – EDN: JOPIJF

RESUME

U. V. Belkin, M. Y. Kuzmin, S. A. Serikov

Comparative analysis of AASIST, RawNet2Spoof and RawNetGatSpoofST architectures for detection of synthesized and deepfake audio recordings

The rapid progress of text-to-speech and voice conversion models has made deepfake audio a realistic threat to biometric systems and media trust. This paper presents a comparative study of three anti-spoofing architectures: AASIST (spectrogram-based graph attention), RawNet2Spoof (raw-signal end-to-end), and the newly proposed RawNetGatSpoofST – a hybrid that combines raw waveform processing with a spectral-temporal graph attention branch, fused via cross-attention. Experiments are conducted on ASVspoof 2019 LA (97k clips) and WaveFake (23k clips). The proposed model achieves an equal error rate (EER) of 0.61% on ASVspoof 2019 LA, outperforming AASIST (0.93%) and RawNet2Spoof (1.18%). Under MP3 compression at 64 kbps, RawNetGatSpoofST retains EER = 1.27%, while AASIST degrades to 4.12%. On cross-generator evaluation with unseen synthesizers, the proposed method yields EER = 2.14% (relative improvement over AASIST: 56%). Inference time on a CPU core is 12 ms per 4-second utterance. RawNetGatSpoofST establishes a new state-of-the-art in audio deepfake detection and is recommended for real-time voice biometrics and content moderation.

РЕЗЮМЕ

У. В. Белкин, М. Ю. Кузьмин, С. А. Сериков

Сравнительный анализ архитектур AASIST, RawNet2Spoof и RawNetGatSpoofST для детекции синтезированной и дипфейк-аудиозаписи

Быстрый прогресс моделей синтеза речи и преобразования голоса сделал дипфейк-аудио реальной угрозой для биометрических систем и доверия к медиа. В работе проведено сравнительное исследование трёх анти-спуфинговых архитектур: AASIST (спектрограммное графовое внимание), RawNet2Spoof (end-to-end на сыром сигнале) и предложенной RawNetGatSpoofST – гибрида, объединяющего обработку сырого сигнала с графовым спектрально-темпоральным вниманием и перекрёстной агрегацией. Эксперименты выполнены на наборах ASVspoof 2019 LA (97 тыс. клипов) и WaveFake (23 тыс. клипов). Предложенная модель достигает EER = 0,61% на ASVspoof 2019 LA, превосходя AASIST (0,93%) и RawNet2Spoof (1,18%). При MP3-сжатии 64 кбит/с RawNetGatSpoofST сохраняет EER = 1,27%, тогда как AASIST деградирует до 4,12%. При кросс-генераторной оценке на не встречавшихся синтезаторах предлагаемый метод даёт EER = 2,14% (относительное улучшение над AASIST – 56%). Время инференса на одном ядре CPU составляет 12 мс на фрагмент длительностью 4 секунды. RawNetGatSpoofST устанавливает новый рекорд качества в детекции аудиодипфейков и рекомендуется для систем голосовой биометрии и модерации в реальном времени.

Белкин Ульян Вячеславович – аспирант, ассистент образовательного офиса инженерной школы, Санкт-Петербургский государственный университет аэрокосмического приборостроения (ГУАП), 190000, Россия, Санкт-Петербург, ул. Большая Морская, д. 67, лит. А. Тел.: +7 (995) 830-15-49. E-mail: i@ubelkin.ru. *Область научных интересов:* системы распознавания дипфейков, роевое управление, робототехника, искусственный интеллект. ORCID: 0009-0003-2433-5649
Сериков Сергей Анатольевич – доктор технических наук, доцент, профессор кафедры электромеханики и робототехники, Санкт-Петербургский государственный университет аэрокосмического приборостроения (ГУАП), 190000, Россия, Санкт-Петербург, ул. Большая Морская, д. 67, лит. А.. Тел.: +7 (931) 541-91-25. E-mail: srkv@inbox.ru *Область научных интересов:* групповая робототехника, нечеткое управление, методы машинного обучения. ORCID: 0000-0002-3668-2459.
Кузьмин Михаил Юрьевич – студент, Санкт-Петербургский государственный университет аэрокосмического приборостроения (ГУАП), 190000, Россия, Санкт-Петербург, ул. Большая Морская, д. 67, лит. А. Тел.: +7 (993) 962-47-48. E-mail: misha.pro17@mail.ru. *Область научных интересов:* системы распознавания аудио дипфейков, искусственный интеллект, методы машинного обучения. ORCID: 0009-0000-8698-6699

Belkin Ulyan Vyacheslavovich – PhD student, Assistant at the Educational Office of the School of Engineering, Saint Petersburg State University of Aerospace Instrumentation (SUAI), 190000, Russia, Saint Petersburg, Bolshaya Morskaya St., 67, Building A. Tel.: +7 (995) 830-15-49. E-mail: i@ubelkin.ru. *Research interests:* deepfake recognition systems, swarm control, robotics, artificial intelligence. ORCID: 0009-0003-2433-5649
Serikov Sergey Anatolyevich – Doctor of Technical Sciences, Associate Professor, Professor at the Department of Electromechanics and Robotics, Saint Petersburg State University of Aerospace Instrumentation (SUAI), 190000, Russia, Saint Petersburg, Bolshaya Morskaya St., 67, Building A. Tel.: +7 (931) 541-91-25. E-mail: srkv@inbox.ru. *Research interests:* group robotics, fuzzy control, machine learning methods. ORCID: 0000-0002-3668-2459
Kuzmin Mikhail Yuryevich – student, Saint Petersburg State University of Aerospace Instrumentation (SUAI), 190000, Russia, Saint Petersburg, Bolshaya Morskaya St., 67, Building A. Tel.: +7 (993) 962-47-48. E-mail: misha.pro17@mail.ru. *Research interests:* audio deepfake detection systems, artificial intelligence, machine learning methods. ORCID: 0009-0000-8698-6699

Статья поступила в редакцию 07.04.2026;
одобрена после рецензирования 12.07.2026; принята к публикации 02.09.2026
The article was submitted 07.04.2026; approved after reviewing 12.07.2026;
accepted for publication 02.09.2026