Алгоритм распознавания синтезированной речи на основе вычисления энтропии аудиосигнала

Д.А. Космынин, Е.К. Григорьев

Санкт-Петербургский государственный университет аэрокосмического приборостроения

Аннотация: В статье предложен алгоритм распознавания синтезированной речи на основе вычисления энтропии аудиосигнала. Актуальность работы обусловлена ростом случаев неправомерного использования синтезированной речи, которая становится практически неотличимой от естественной. Результаты показали, что энтропия синтезированной речи значительно выше, а алгоритм устойчив к потерям данных. Преимуществами алгоритма являются простота интерпретации и невысокая вычислительная сложность. Эксперименты проведены на датасете CMU ARCTIC с использованием модели XTTS v.2. Предложенный алгоритм позволяет принять решение о наличии синтезированной речи без необходимости применения сложных методов спектрального анализа и машинного обучения.

Ключевые слова: синтезированная речь, спуффинг, энтропия Шэннона, распознавание речи.

Введение

Развитие технологий синтеза речи, а также генеративных моделей, привело к тому, что синтезированная речь стала практически неотличима от естественной человеческой речи. Параллельно с этим, наблюдается рост случаев неправомерного использование подобных технологий [1,2].

Современные подходы к распознаванию синтезированной речи в большинстве случаев основаны на анализе специфических акустических, спектральных или лингвистических закономерностей, которые были оставлены алгоритмами синтеза речи.

В области анализа спектральных характеристик синтезированной речи известны методы с использованием мел-кепстральных характеристик [3], и спектрограмм [4], которые в настоящее время в основном используются в комбинации с нейронными сетями [4,5]. В настоящее время наиболее

популярные и эффективные подходы к распознаванию синтезированной речи связанны с использованием нейронных сетей. Использование различных архитектур нейронных сетей для распознавания синтезированной речи приведено в обзоре [6]. Отдельно следует выделить подход с использованием лингвистических и просодических особенностей, поскольку синтезированная речь может нарушать естественные патттерны ударения, ритма или эмоциональной окраски диктора [7]. Также представляется перспективным описанный в [8] способ выявления синтезированной речи на основе поддиапазонных частот, которые в связи с Q-константными кепстральными коэффициентами существенно лучше улавливают спектральные закономерности, уникальные для реальной речи.

Однако, несмотря на значительное количество подходов к распознаванию синтезированной речи, дальнейший поиск новых методов остается актуальной задачей, поскольку анализ существующих решений, например перечисленных в обзоре [6], показал, что существующие решения зачастую демонстрируют высокую эффективность только на конкретных типах синтезаторов или наборах данных, что ограничивает их применимость в реальных условиях. В этой связи дальнейший поиск методов распознавания синтезированной речи является необходимым для обеспечения защиты от новых угроз, а также поддержания доверия к существующим биометрическим системам. В связи с вышесказанным, целью настоящей работы является предложение алгоритма распознавания синтезированной речи на основе статистических характеристик аудиофайлов, а именно энтропии.

Структурная схема и описание основных этапов работы алгоритма

Предлагаемый алгоритм базируется на идеях, заложенных в работах [8-9]. В основе его работы находится понятие энтропии – меры хаотичности и неопределенности полезной составляющей, содержащейся в сигнале.

Структурная схема алгоритма представлена на рисунке 1.

Рис. 1. – Структурная схема алгоритма распознавания синтезированной речи

Поясним основные этапы работы алгоритма. На первом этапе аудиозапись проходит предобработку, включающую в себя RMS-нормализацию громкости [5], фильтрацию неинформативной части сигнала, а также выравнивание длительности образцов подлинной и синтезированной речи. После предобработки на втором и третьем этапе вычисляются значения энтропии для подлинной и синтезированной речи, по методу, описанному нами в работе [9]. Для работы алгоритма необходимо выполнить следующие действия.

1. Исходный сигнал разбивается на i равных сегментов;

2. Для каждого i-го сегмента строится огибающая минимумов и максимумов, вычисляются средние экстремумы;

3. Отрезок разбивается n равных частей, называемых элементарными сообщениями;

4. Множество элементарных сообщений дополняется необходимым количеством отрезков k оси Y так, чтобы значения глобального минимума и максимума входили в крайние отрезки, тогда общее количество элементарных сообщений равно $n+k$;

5. Вычисляется итоговое значение энтропии $H(X)$ по формуле:

$$H\left(X\right)=-\sum_{i=1}^{n}p_i\log_2\left(p_i\right)$$

Числа n и k, представляющие собой число разбиений интервалов, выбираются произвольно. Экспериментально было выявлено, что сбалансированное число таких разбиений n=100000, k=200000.

Экспериментальные исследования

Для проведения эксперимента был использован датасет CMU ARCTIC, содержащий в себе образцы фраз реальной речи, принадлежащих различным англоязычным говорящим. На основе аудиозаписей из приведенного датасета были сгенерированы образцы синтезированной речи при помощи Text-to-speech модели XTTS v.2. Отметим, что эксперименты были проведены с полным датасетом, но в рамках демонстрации мы ограничились фразами, приведёнными в таблице 1.

Таблица № 1

Исходные фразы длительностью 5 секунд

Номер фразы	Содержимое фразы
1	Offered of the danger trail, Philip steals, et cetera.
2	Not at this particular case, Tom apologised with more.
3	For the 20th time that evening the two men shook hands.
4	He was a head shorter than his companion of almost delicate physique.

В качестве примера на рис. 2-5 приведены результаты работы алгоритма для фраз 1-4 из таблицы 1 соответственно.

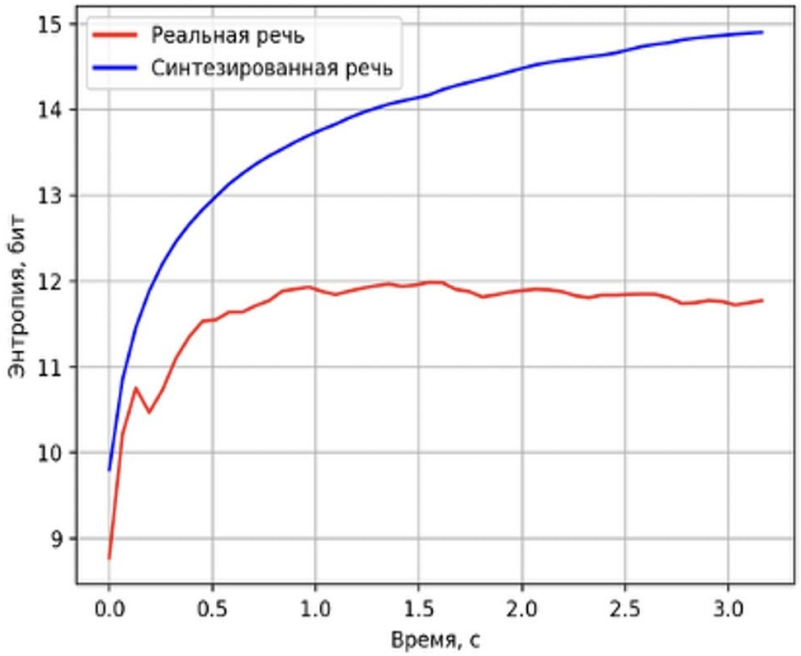

Рис. 2. – График зависимости энтропии от времени для фразы 1

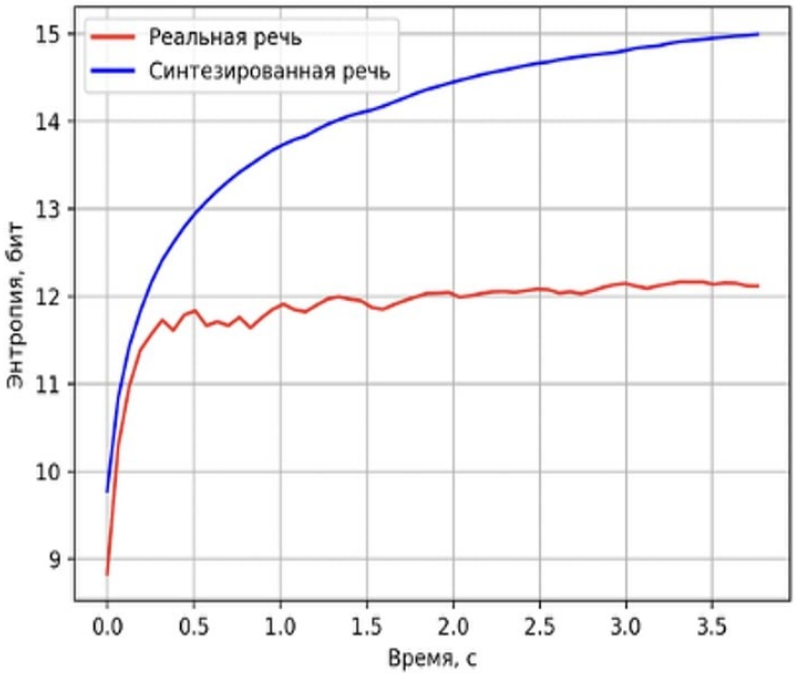

Рис. 3. – График зависимости энтропии от времени для фразы 2

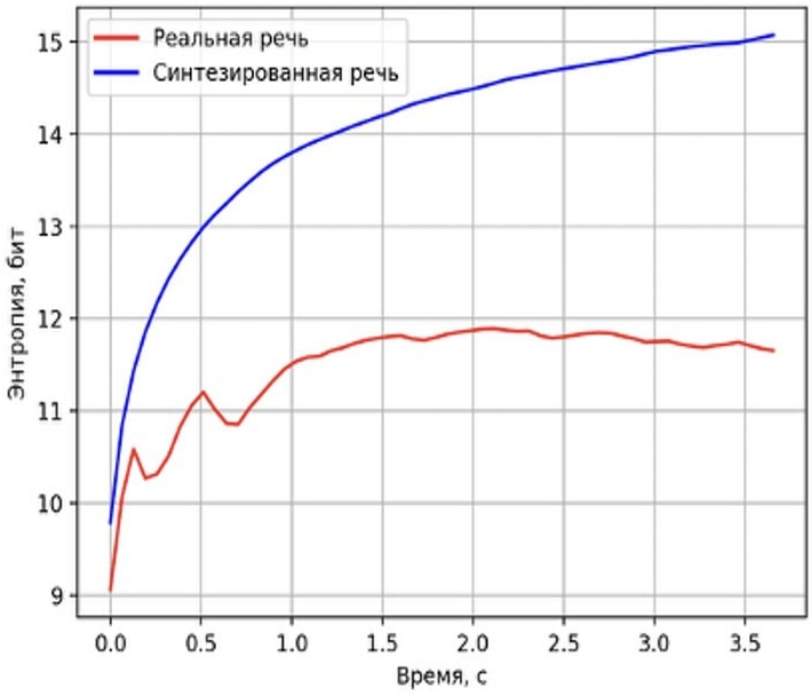

Рис. 4. – График зависимости энтропии от времени для фразы 3

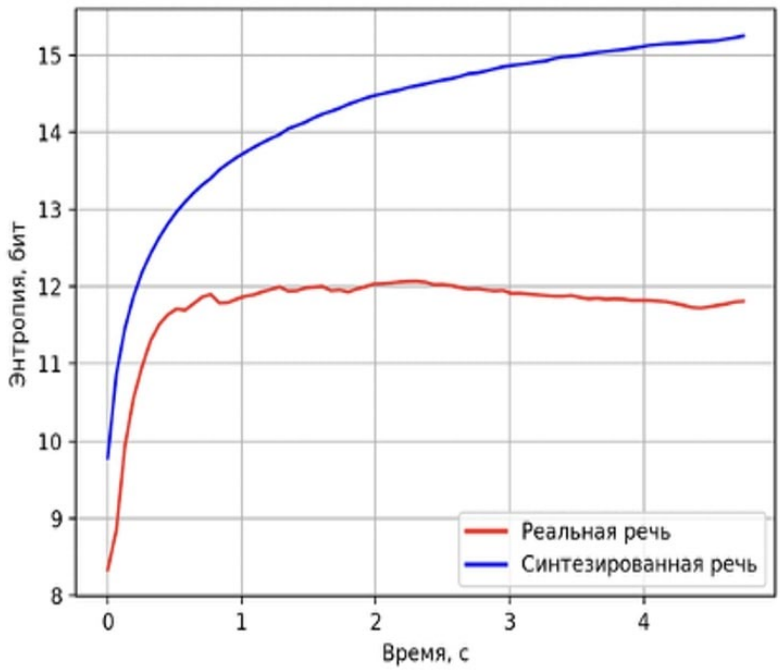

Рис. 5. – График зависимости энтропии от времени для фразы 4

Нетрудно заметить, что при помощи алгоритма можно осуществить распознавание синтезированной речи, поскольку значение ее энтропии во всех случаях будет значительно выше.

Преимуществом предлагаемого алгоритма является возможность распознавания синтезированной речи по скалярному значению вычисляемой энтропии, в отличии от методов и алгоритмов, которые в своей основе используют анализ спектральных различий аудиосигналов. В качестве примера представлены рис. 6-7, на которых показаны спектрограммы для подлинной и синтезированной речи фразы 3 из табл.1.

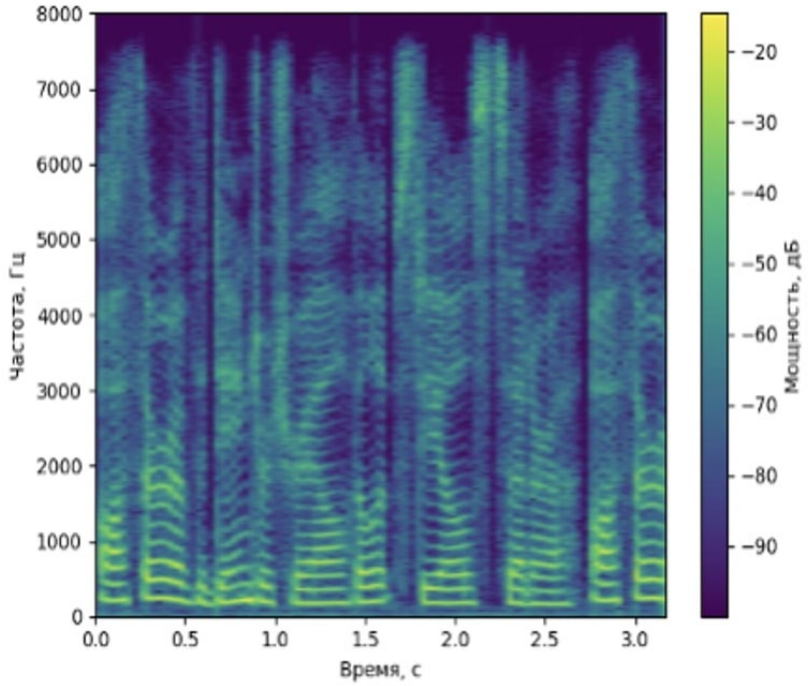

Рис. 6. – Спектрограмма подлинной речи для фразы 3

Рисунки позволяют заметить некоторые спектральные различия двух аудиосигналов, однако эти различия являются слабо уловимыми для человеческого восприятия, что создает затруднения при формировании решения и требует привлечения нейронных сетей [4,5].

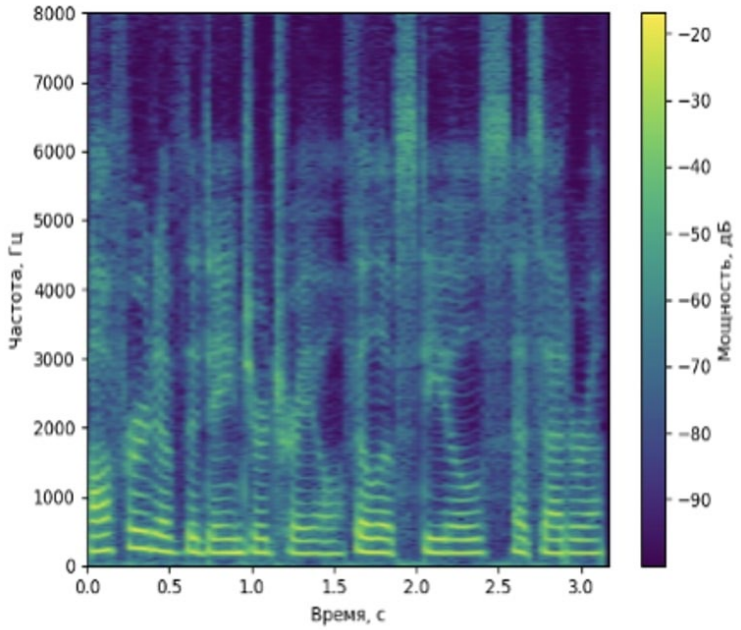

Рис. 7. – Спектрограмма синтезированной речи для фразы 3

Дополнительно было решено протестировать работу алгоритма в случае потери пакетов при передаче данных, результаты которого приведены в табл. 2.

Таблица № 2

Энтропия пятисекундных фраз при потерях в данных

Энтропия	Фраза 1	Фраза 2	Фраза 3	Фраза 4
Подлинная речь	11.782	12.108	11.651	11.822
Подлинная речь, с потерями	11.759	12.113	11.675	11.811
Синтезированная речь	14.858	14.995	15.070	15.218
Синтезированная речь, с потерями	14.842	14.987	15.027	15.202

Для проведения эксперимента производился расчёт энтропии исходного аудиосигнала, а затем удалялся некоторый случайный сегмент сигнала, не превышающий 10% от длины аудиофайла, после чего энтропия

рассчитывалась повторно. Нетрудно заметить, что предлагаемый алгоритм мало чувствителен к потерям данных.

В виду отсутствия открытых данных о вычислительных затратах алгоритмов, которые используются для решения аналогичной задачи было решено сравнить предлагаемое решение с решением из работы [5]. Сразу отметим, что предлагаемое в работе [5] решение более комплексное и включает в себя стадии предобработки с вычислением кепстральных коэффициентов и классификации на основе методов машинного обучения, поэтому было решено сравнить вычислительную сложность нашего алгоритма и стадию предобработки из работы [5].

Таблица №3

Алгоритмическая сложность предлагаемого решения и решения из работы [5]

Вычислительная сложность	По времени	По памяти
Предлагаемое решение	O(n) + O(N) + O(N*log(N))	O(N) + O(e)
Решение из работы [5]	O(N) + O(s) + O(N log(N)) + O(M log(M)) + O(z) + O(M) +O(KT log(KT)) + O(N K) + O(KT)	O(N) + O(e) + O(z) + O(M) + O(KT)

где буквами обозначены следующие величины:

- для предлагаемого решения: N – количество дискретных значений сигнала, n – количество сегментов в разбиении, e – количество обнаруженных экстремумов для сегмента;

- для решения из работы [5]: N - число дискретных значений сигнала, s - количество разбиений исходного сигнала на отрезки равных частей, M - количество коэффициентов, полученных при расчете энергий мел-

фильтров, z – количество частотных компонент сигнала, K - количество частотных компонент в Q-преобразовании, где T - количество временных окон, e – количество обнаруженных экстремумов для сегмента.

Заключение

В статье предложен алгоритм распознавания синтезированной речи на основе вычисления энтропии аудиосигнала. Распознавание можно провести по вычисляемому алгоритмом значению энтропии, которое для синтезированной речи будет всегда значительно выше по сравнению с подлинной речью.

Алгоритм позволяет провести распознавание по скалярному значению вычисляемой энтропии что позволяет пропустить на этапе принятия решения визуальный анализ или привлечение методов машинного обучения, в отличии от методов и алгоритмов, использующих спектральные различия аудиосигналов. Предлагаемый алгоритм в результате экспериментов показал робастность к потерям данных при передаче по коммуникационному каналу.

Разработанный подход и алгоритм на его основе может быть при необходимости расширен до бинарного классификатора аудиосигналов, использующий энтропию и распределение вероятностей как свойства обучающей выборки модели машинного обучения.

Литература

1. Евсюков М. В., Путято М. М., Макарян А. С. Исследование различимости подлинного и синтезированного голоса дикторов. Вопросы кибербезопасности. 2024. № 2(60). С. 44-52. DOI: 10.21681/2311-3456-2024-2-44-52.

2. Пономарёв К.Г., Верещагиина Е.А. Применение сверточных нейронных сетей и алгоритмов глубокого обучения для прогнозирования и идентификации голосовых дипфейков // Инженерный вестник Дона, 2025, №2 URL: ivdon.ru/ru/ magazine/archive/n2y2025/9859/.

3. Дашинимаева В.Ц., Боршевников А.Е. Методы обнаружения ложных голосовых сигналов // Инженерный вестник Дона, 2025, №8 URL: ivdon.ru/ru/magazine/archive/n8y2025/10326/.

4. Nautsch A. ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech. IEEE Transactions on Biometrics, Behavior, and Identity Science. vol. 3. no. 2. pp. 252-265. April 2021. DOI: 10.1109/TBIOM.2021.3059479.

5. Муртазин Р. А, Кузнецов А. Ю., Федоров Е. А., Гарипов И. М., Холоденина А. В., Балданова Ю. Б., Воробьева А.А. Алгоритм выявления синтезированного голоса на основе кепстральных коэффициентов и сверточной нейронной сети. Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 4. С. 545-552. DOI: 10.17586/2226-1494-2021-21-4-545-552.

6. Mubarak R., Alsboui T., Alshaikh O., Inuwa-Dutse I., Khan S. and Parkinson S. A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. IEEE Access. vol. 11. pp. 144497-144529. 2023. DOI: 10.1109/ACCESS.2023.3344653.

7. Ненашева Ю. А. Просодический фактор в синтезировании речи. Когнитивные исследования языка. 2024. № 2-2(58). С. 630-634.

8. Yang, R. K. Das, H. Li, «Significance of Subband Features for Synthetic Speech Detection», IEEE Trans. Inf. Forensics Security, Vol. 15, pp. 2160–2170, 2020, DOI: 10.1109/TIFS.2019.2956589.

9. Суворова В. А., Шкарапута А. П. Разработка и возможности применения алгоритма вычисления энтропии звуковой волны. Вестник Пермского университета. Математика. Механика. Информатика. 2016. № 1(32). С. 29-33.

10. Космынин Д. А., Григорьев Е. К. Метод распознавания синтезированной речи на основе вычисления энтропии аудиосигнала.

Прикладной искусственный интеллект: перспективы и риски: Сборник докладов Международной научной конференции, Санкт-Петербург, 17 октября 2024 года. Санкт-Петербург: Санкт-Петербургский государственный университет аэрокосмического приборостроения. 2024 С. 184-189.

References

1. Evsjukov M. V., Putjato M. M., Makarjan A. S. Voprosy kiberbezopasnosti. 2024. № 2(60). pp. 44–52. DOI: 10.21681/2311-3456-2024-2-44-52.

2. Ponomarev K.G., Vereshchagiina E.A. Inzhenernyj vestnik Dona, 2025, №2. URL: ivdon.ru/ magazine/archive/n2y2025/9859/.

3. Dashinimayeva V.TS., Borshevnikov A.E. Inzhenernyj vestnik Dona, 2009, №1. URL: ivdon.ru/ magazine/archive/n8y2025/10326/.

4. Nautsch A. IEEE Transactions on Biometrics, Behavior, and Identity Science. Vol. 3. №2. pp. 252-265. April 2021. DOI: 10.1109/TBIOM.2021.3059479.

5. Murtazin R. A., Kuznetsov A. J., Fedorov E. A., Garipov I. M., Kholodenina A. V., Baldanova J. B., Vorobyeva A. A. Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki. 2021. T. 21. № 4. pp. 545–552. DOI: 10.17586/2226-1494-2021-21-4-545-552.

6. Mubarak R., Alsboui T., Alshaikh O., Inuwa-Dutse I., Khan S., Parkinson S. IEEE Access. Vol. 11. pp. 144497-144529. 2023. DOI: 10.1109/ACCESS.2023.3344653.

7. Nenasheva J. A. Kognitivnye issledovanija jazyka. 2024. № 2-2(58). pp. 630–634.

8. Yang, R. K. Das, H. Li. IEEE Trans. Inf. Forensics Security, Vol. 15, pp. 2160–2170, 2020, DOI: 10.1109/TIFS.2019.2956589.

9. Suvorova V. A., Shkaraputa A. P. Vestnik Permskogo universiteta. Matematika. Mehanika. Informatika. 2016. № 1(32). pp. 29–33.

10. Kosmynin D. A., Grigor’ev E. K. Prikladnoj iskusstvennyj intellekt: perspektivy i riski: Sbornik dokladov Mezhdunarodnoj nauchnoj konferencii, Sankt-Peterburg, 17 oktjabrja 2024 goda. Sankt-Peterburg: Sankt-Peterburgskij gosudarstvennyj universitet ajerokosmicheskogo priborostroenija, 2024. pp. 184–189.

Дата поступления: 1.08.2025

Дата публикации: 25.09.2025