УДК 004.932.2
DOI 10.17513/snt.40733

ПРИМЕНИМОСТЬ ДИФФУЗИОННЫХ МОДЕЛЕЙ ДЛЯ ШУМОПОДАВЛЕНИЯ КЛИНИЧЕСКОЙ РЕЧИ: ЭКСПЕРИМЕНТАЛЬНЫЙ АНАЛИЗ И ОГРАНИЧЕНИЯ

[1, 2]Нуреев А. Р., [3] Староверова Н. А. ORCID ID 0000-0002-5524-1325

[1]Общество с ограниченной ответственностью «ПЭСТ», Казань, Российская Федерация;
[2]Нижнекамский химико-технологический институт (филиал) федерального государственного бюджетного образовательного учреждения высшего образования «Казанский национальный исследовательский технологический университет», Нижнекамск, Российская Федерация;
[3] Федеральное государственное бюджетное образовательное учреждение высшего образования «Казанский национальный исследовательский технологический университет», Казань, Российская Федерация, e-mail: nata-staroverova@yandex.ru

Работа посвящена экспериментальной оценке применимости диффузионных вероятностных моделей для шумоподавления клинической речи. Исследование носит характер обмена опытом и направлено на сопоставление диффузионного подхода с классическим алгоритмом OM-LSA в контролируемых условиях моделирования шумов. Проведен анализ и формализация компонентной структуры нестационарных клинических шумов (стационарный фон, импульсные помехи, реверберация). На основе этого анализа выдвинуто и проверено центральное утверждение о принципиальной ограниченности классических линейных методов в данных условиях. В эксперименте использован синтетический корпус медицинской речи объемом 4,2 ч (2350 фрагментов, 18 дикторов, 16 кГц) с наложением стационарных, узкополосных и импульсных помех при SNR от +5 до -5 дБ. Для каждого уровня SNR генерировались три независимые реализации шума. Сравнение проводилось по метрикам PESQ, STOI, относительному искажению формантных частот (ΔF1, ΔF2) и Word Error Rate (WER) системы распознавания речи на базе wav2vec 2.0. Показано, что диффузионная модель демонстрирует лучшее сохранение формантной структуры и более существенное снижение WER по сравнению с OM-LSA при сопоставимых условиях. Одновременно выявлены вычислительные и методические ограничения подхода.

Ключевые слова: диффузионные модели, шумоподавление речи, клиническая акустическая среда, медицинская речь, нестационарный шум, теоретическое обоснование, формантный анализ, робастное распознавание речи

APPLICABILITY OF DIFFUSION MODELS FOR NOISE REDUCTION OF CLINICAL SPEECH: EXPERIMENTAL ANALYSIS AND LIMITATIONS

[1, 2]Nureev A. R., [3]Staroverova N. A. ORCID ID 0000-0002-5524-1325

[1]PEST Limited Liability Company, Kazan, Russian Federation;
[2]Nizhnekamsk Chemical and Technological Institute (branch) of the Federal State Budgetary Educational Institution of Higher Education "Kazan National Research Technological University", Nizhnekamsk, Russian Federation;
[3]Federal State Budgetary Educational Institution of Higher Education "Kazan National Research Technological University", Kazan, Russian Federation, e-mail: nata-staroverova@yandex.ru

The work is devoted to the experimental assessment of the applicability of diffusion probabilistic models for noise reduction in clinical speech. The study is an exchange of experience and aims to compare the diffusion approach with the classical OM-LSA algorithm under controlled noise modeling conditions. The analysis and formalization of the component structure of non-stationary clinical noises (stationary background, impulse interference, and reverberation) have been conducted. Based on this analysis, a central claim has been made and verified regarding the fundamental limitations of classical linear methods under these conditions. The experiment used a synthetic corpus of medical speech with a volume of 4.2 hours (2350 fragments, 18 speakers, 16 kHz) with stationary, narrow-band, and impulse interference at SNR from +5 to –5 dB. Three independent noise realizations were generated for each SNR level. The comparison was carried out using the PESQ, STOI, relative distortion of formant frequencies (ΔF1, ΔF2), and Word Error Rate (WER) metrics of the wav2vec 2.0 speech recognition system. It was shown that the diffusion model demonstrates better preservation of the formant structure and a more significant reduction in WER compared to OM-LSA under comparable conditions. At the same time, the computational and methodological limitations of the approach were identified.

Keywords: diffusion probabilistic models, speech enhancement, clinical acoustic environment, medical speech, non-stationary noise, theoretical justification, formant analysis, robust automatic speech recognition

Введение

Диффузионные вероятностные модели утвердились как мощный инструмент генерации и восстановления данных, демонстрируя конкурентоспособные результаты в обработке изображений и аудио [1, 2]. Вместе с тем диффузионные модели обладают рядом существенных ограничений: высокой вычислительной сложностью обратного процесса, требовательностью к объему обучающих данных, стохастическим характером восстановления и нестабильностью при экстремально низком SNR. Эти особенности требуют отдельной оценки применимости метода в условиях клинической акустической среды.

Их теоретический потенциал для задач шумоподавления речи, трактуемого как восстановление сигнала из сложного апостериорного распределения, является очевидным [2]. Однако прямое перенесение этих результатов в узкую и критически важную область клинической речи сталкивается с существенным пробелом. Существует разрыв между абстрактными теоретическими возможностями диффузионных моделей и их обоснованным, методически выверенным применением в условиях конкретной акустической среды медицинских учреждений. Этот разрыв обусловлен недостаточным учетом в существующих исследованиях уникальной специфики клинических шумов и отсутствием целевых экспериментальных доказательств, связывающих свойства моделей с требованиями медицинской практики [3, 4].

Клиническая среда характеризуется не просто низким отношением сигнал/шум (SNR), а сложной аддитивно-компонентной структурой помех (стационарный фон, импульсные сигналы, реверберация), обладающих высокой нестационарностью и спектральным перекрытием с речевым диапазоном. Традиционные методы, доминирующие в практических реализациях, основаны на линейных предположениях, заведомо неадекватных в таких условиях.

Цель исследования – экспериментальная проверка гипотезы о том, что в условиях нестационарных и спектрально перекрывающихся клинических шумов вероятностные диффузионные модели обеспечивают более высокое качество восстановления речевого сигнала (в частности, сохранение формантной структуры) по сравнению с линейными методами спектрального вычитания. Для достижения этой цели проводим теоретический анализ специфики клинической акустической среды и формализуем ее в виде модели шума, а затем в контролируемом эксперименте сопоставляем два подхода. Проверяемое утверждение формулируется следующим образом: «При наличии нестационарных и спектрально перекрывающихся клинических шумов методы шумоподавления, основанные на линейной фильтрации и спектральном вычитании, не обеспечивают одновременного подавления помех и сохранения клинически значимых акустических признаков речи, что требует перехода к нелинейным вероятностным методам».

Материал и методы исследования

Методология исследования была построена как двухэтапный процесс, объединяющий теоретический анализ и целенаправленную экспериментальную проверку.

Теоретический анализ и формализация специфики среды

На основе обзора литературы и анализа реальных записей [4, 5] была разработана формальная модель клинического шума:

$$n(t) = n_s(t) + n_i(t) + n_r(t),$$

где n_s – стационарный фон (вентиляция), n_i – импульсные помехи (сигналы тревог), n_r – реверберационная составляющая.

Эта модель легла в основу синтеза тестовых данных и позволила сформулировать проверяемое утверждение.

Экспериментальная проверка как элемент обоснования

Для верификации утверждения и обоснования адекватности диффузионного подхода был создан контролируемый эксперимент.

В качестве корпуса данных использовался синтетический набор на основе чистых медицинских диалогов [6, 7], к которым аддитивно добавлялись смоделированные компоненты $n_s(t)$ (узкополосный шум в области формант) и $n_i(t)$ (импульсные последовательности). SNR варьировался от +5 до -5 дБ. Общий объем чистой речи составил 4,2 ч (2350 фрагментов, 18 дикторов, частота дискретизации 16 кГц). Средняя длительность одного фрагмента – 6,4 с. Для каждого уровня SNR (+5, 0, –5 дБ) генерировались три независимые реализации шума. Общее количество зашумленных примеров составило 7050.

В качестве репрезентанта класса методов, чья неадекватность постулируется утверждением, выбран алгоритм OM-LSA [8], основанный на модифицированном логарифмическом спектральном вычитании и предполагающий линейную модель аддитивной помехи.

В качестве альтернативного подхода использована условная диффузионная модель (DDPM) с архитектурой U-Net, реализующая восстановление чистого речевого сигнала при условии наблюдаемого зашумленного сигнала. В рамках данной работы под условной моделью понимается вероятностная модель, аппроксимирующая распределение:

$$p(x \mid y),$$

где x – чистая речь, y – соответствующий зашумленный сигнал.

Обратный диффузионный процесс параметризуется в виде

$$p_\theta(x_{t-1} \mid x_t, y),$$

что означает учет акустического контекста зашумленного наблюдения на каждом шаге восстановления. Кондиционирование реализовано путем подачи признаков зашумленного сигнала в сеть U-Net совместно с текущим диффузионным состоянием.

Число шагов диффузии составляло 1000. Обучение проводилось в течение 150 эпох при размере батча 16. Время обучения составило около 42 ч (GPU RTX 3090).

Оценка качества восстановления речевого сигнала проводилась по многоуровневой системе показателей, включающей интегральные акустические метрики, анализ сохранности фонетически значимых признаков и влияние обработки на точность автоматического распознавания речи.

Интегральные показатели качества сигнала

PESQ (Perceptual Evaluation of Speech Quality) [9] – перцептивная оценка качества речи.

Данная метрика моделирует особенности слухового восприятия человека и сравнивает обработанный сигнал с эталонным чистым сигналом. Значения PESQ находятся в диапазоне от 1 до 4,5, где большие значения соответствуют лучшему субъективному качеству.

STOI (Short-Time Objective Intelligibility) [10] – объективная кратковременная мера разборчивости речи.

Метрика основана на корреляции временно-частотных представлений чистого и обработанного сигнала и оценивает степень сохранения разборчивости. Значения лежат в диапазоне от 0 до 1 (или в процентах от 0 до 100 %).

Сохранение клинически значимых признаков

В качестве ключевого критерия использовалось среднее относительное отклонение формантных частот первого (F1) и второго (F2) формант:

$$\Delta F = \frac{\left|F_{\text{обраб}} - F_{\text{чист}}\right|}{F_{\text{чист}}} \cdot 100\ \%.$$

Порог в 10 % принят в соответствии с психоакустическими исследованиями как граница, после которой искажения могут приводить к изменению фонематической идентификации гласных [11].

Влияние на точность автоматического распознавания речи

Использовался показатель WER (Word Error Rate) – относительная частота ошибок распознавания слов.

Показатель вычисляется как [12]:

$$WER = \frac{S + D + I}{N},$$

где S – число замен,

D – число пропусков,

I – число вставок,

N – общее число слов в эталонной транскрипции.

Распознавание выполнялось с использованием модели wav2vec 2.0 [13].

Статистическая обработка результатов

Для каждого уровня отношения сигнал/шум использовались три независимые реализации помехи. Итоговые значения метрик представлены как средние по трем экспериментам [14]. Стандартное отклонение WER для диффузионной модели составило до 1,3 %, что отражает вероятностный характер обратного диффузионного процесса.

Результаты исследования и их обсуждение

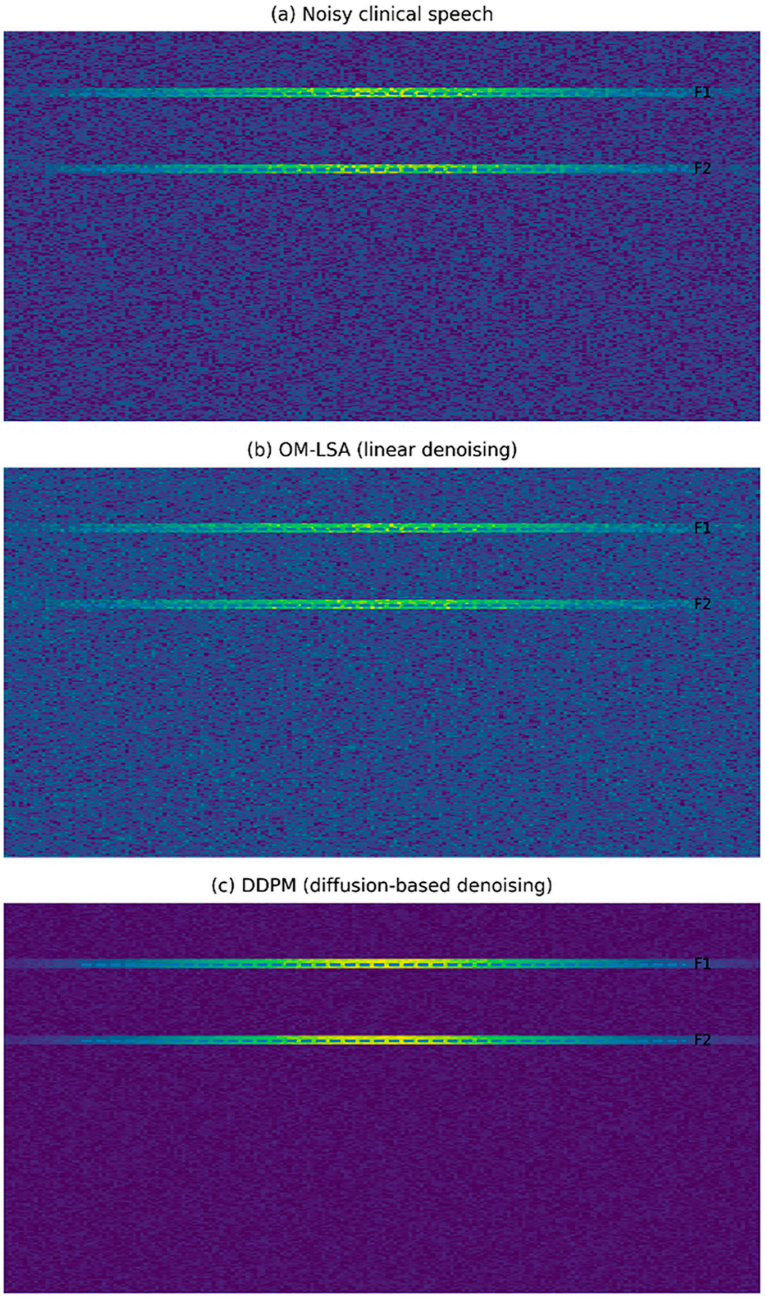
Полученные результаты (таблица, рисунок) служат прямым экспериментальным доказательством, заполняющим целевой пробел между теорией и практикой применения диффузионных моделей в клинике.

Полученные экспериментальные результаты однозначно подтверждают сформулированное во введении утверждение о принципиальной ограниченности линейных методов шумоподавления в условиях клинической акустической среды. Алгоритм OM-LSA, представляющий класс методов спектрального вычитания, демонстрирует ожидаемое улучшение интегральных энергетических метрик качества речи (PESQ и STOI). Однако данное улучшение носит преимущественно формальный характер и не сопровождается сохранением клинически значимых акустических признаков.

Результаты обработки при SNR = 0 дБ

Метод (Парадигма)	PESQ	STOI, %	ΔF1, %	ΔF2, %	WER, %	ΔWER отн. зашумл.
Зашумленный сигнал	1,42	62,3	–	–	38,7	–
OM-LSA (Линейная)	2,01	71,5	18,4	21,7	34,9	−3,8
DDPM (Вероятностная)	2,34	78,9	6,2	7,1	24,6	−14,1

Примечание: составлена авторами на основе полученных данных в ходе исследования



Сравнение спектрограмм: (a) чистый сигнал с узкополосной помехой (отмечена стрелкой); (b) после обработки OM-LSA (видны искажения формант, отмечены кругами); (c) после обработки DDPM (форманты восстановлены, помеха подавлена)
Примечание: составлен авторами по результатам данного исследования

Анализ относительного искажения формантных частот показывает, что после обработки OM-LSA средние значения ΔF1 и ΔF2 существенно превышают критический порог в 10 %, принятый в настоящей работе как граница фонематической целостности. Данный факт указывает на систематическое смещение формантных областей, обусловленное спектральным перекрытием клинических шумов с речевыми компонентами. Поскольку алгоритм OM-LSA не обладает механизмом разделения шума и речи в условиях совпадения их спектральных характеристик, подавление помех осуществляется за счет подавления самих речевых формант.

Ключевым следствием этого является несоответствие между улучшением объективных метрик качества сигнала и практической эффективностью для downstream-задачи автоматического распознавания речи. Несмотря на рост PESQ и STOI, снижение WER после применения OM-LSA носит незначительный характер. Это свидетельствует о том, что искажение формантной структуры нивелирует потенциальную пользу от подавления шума и приводит к утрате акустических признаков, критически важных для корректной фонематической декодировки.

Таким образом, экспериментально подтверждается, что линейная парадигма шумоподавления в условиях нестационарных и спектрально перекрывающихся клинических помех не обеспечивает одновременного выполнения двух ключевых требований: эффективного подавления шума и сохранения клинически значимой структуры речевого сигнала. Данный результат следует рассматривать не как частный недостаток конкретного алгоритма, а как следствие фундаментальных предположений линейной модели, нарушающихся в медицинской акустической среде.

В противоположность линейным методам, условная диффузионная модель демонстрирует качественно иной характер восстановления речевого сигнала, согласующийся с теоретическими предпосылками вероятностного подхода. Полученные экспериментальные данные показывают, что применение DDPM приводит не только к улучшению энергетических метрик качества речи, но, что принципиально важно, к сохранению формантной структуры в пределах допустимых отклонений.

Значения ΔF1 и ΔF2 после обработки диффузионной моделью стабильно остаются ниже критического порога, что указывает на сохранение фонематической целостности речевого сигнала [11, 15]. Это свидетельствует о том, что обратный диффузионный процесс не осуществляет локальное подавление спектральных компонент, а выполняет восстановление сигнала в пространстве допустимых речевых реализаций, определяемом априорным распределением речи. В результате шумовые компоненты устраняются без разрушения формантных областей, даже при их спектральном перекрытии с помехами.

Данный эффект находит прямое отражение в показателях точности автоматического распознавания речи. Существенное снижение WER, по сравнению как с зашумленным сигналом, так и с результатом линейной обработки, указывает на то, что диффузионное шумоподавление формирует акустические представления, более согласованные с требованиями современных ASR-систем [12, 13]. Таким образом, улучшение качества сигнала в данном случае имеет не формальный, а функционально значимый характер.

С методологической точки зрения полученные результаты служат экспериментальным подтверждением того, что вероятностный характер обратного диффузионного процесса позволяет эффективно аппроксимировать сложное апостериорное распределение речевого сигнала в условиях неопределенности, создаваемой клинической акустической средой. Именно это свойство заполняет выявленный разрыв между теоретическими возможностями диффузионных моделей и их практической применимостью в медицинском контексте.

Следовательно, специфика клинической акустической среды – высокая нестационарность, импульсный характер помех и спектральное перекрытие с речью – делает диффузионные модели не просто альтернативным инструментом шумоподавления, а теоретически обоснованным и экспериментально подтвержденным решением для задач обработки медицинской речи.

Важно отметить, что диффузионная модель, будучи вероятностной, демонстрирует некоторую вариабельность результатов. Стандартное отклонение WER при обработке трех независимых реализаций шума составило до 1,3 %, что отражает стохастическую природу обратного процесса. Тем не менее полученный разброс существенно меньше достигаемого снижения WER (14,1 % относительно зашумленного сигнала), поэтому на практике модель можно считать стабильной.

При ухудшении отношения сигнал/шум до -5 дБ эффективность DDPM снижается: наблюдались случаи неполного подавления импульсных помех, что проявилось в росте WER до 32 % и увеличении искажений

формант. Это указывает на границы применимости метода в условиях экстремально низкого SNR и требует дальнейших исследований, например использования предобученных акустических моделей или комбинированных подходов.

Ограничения исследования

1. Эксперимент выполнен на синтетически зашумленном корпусе.

2. Тестирование в реальной клинической среде не проводилось.

3. Время инференса диффузионной модели превышает время работы OM-LSA примерно в 28 раз.

4. При SNR -5 дБ наблюдались случаи неполного подавления импульсных помех.

5. Модель требует предварительного обучения на размеченных данных.

Заключение

Проведенный эксперимент показывает перспективность использования диффузионных моделей для шумоподавления клинической речи в условиях контролируемого моделирования шумов. Диффузионный подход демонстрирует лучшее сохранение формантной структуры и более выраженное снижение WER по сравнению с алгоритмом OM-LSA. Одновременно выявлены ограничения, связанные с вычислительной сложностью и стохастичностью восстановления. Полученные результаты не претендуют на окончательное доказательство превосходства метода и требуют дальнейшей проверки в реальной клинической среде, а также исследования способов снижения вычислительной нагрузки и повышения робастности при сверхнизких SNR.

Список литературы

1. Croitoru F. A., Hondru V., Ionescu R. T., Shah M. Diffusion models in vision: A survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. Vol. 45. Is. 9. P. 10850–10869. DOI: 10.1109/TPAMI.2023.3261988.

2. Lu Y. J., Wang Z. Q., Watanabe S., Richard A., Yu C., Tsao Y. Conditional diffusion probabilistic model for speech enhancement // 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. P. 7402–7406. DOI: 10.1109/ICASSP43922.2022.9746901.

3. Sarker A., Zhang R., Wang Y., Xiao Y., Das S., Schutte D., Oniani D., Xie Q., Xu H. Natural language processing for digital health in the era of large language models // Yearbook of Medical Informatics. 2024. Vol. 33. Is. 1. P. 229–240. DOI: 10.1055/s-0044-1800750.

4. Zhang L., Cheng W., Zhao M., Tang H. Effect of acoustic environment in wards on postoperative rehabilitation in patients with oral cancer: A retrospective study // Noise and Health. 2024. Vol. 26. Is. 121. P. 148–152. DOI: 10.4103/nah.nah_34_24.

5. Лебедев Г. С., Шадеркин И. А., Лебедева Н. А. Модифицируемые факторы среды помещения: влияние на здоровье человека и цифровой мониторинг. Аналитический обзор // Журнал телемедицины и электронного здравоохранения. 2023. Т. 9. № 1. С. 21–48. DOI: 10.29188/2712-9217-2023-9-1-21-48.

6. Johnson A. E. W., Pollard T. J., Shen L. et al. MIMIC-III, a freely accessible critical care database // Scientific Data. 2016. Vol. 3. Is. 1. P. 1–9. DOI: 10.1038/sdata.2016.35.

7. Czyżewski A. et al. A comprehensive Polish medical speech dataset for enhancing automatic medical dictation // Scientific Data. 2025. Vol. 12. Is. 1. P. 1436. DOI: 10.1038/s41597-025-05776-1.

8. Cohen I., Berdugo B. Speech enhancement for non-stationary noise environments // Signal Processing. 2001. Vol. 81. Is. 11. P. 2403–2418. DOI: 10.1016/S0165-1684(01)00128-1.

9. Rix A. W., Beerends J. G., Hollier M. P., Hekstra A. P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs // 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2001. Vol. 2. P. 749–752. DOI: 10.1109/ICASSP.2001.941023.

10. Taal C. H., Hendriks R. C., Heusdens R., Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech // IEEE Transactions on Audio, Speech, and Language Processing. 2011. Vol. 19. Is. 7. P. 2125–2136. DOI: 10.1109/TASL.2011.2114881.

11. Buder E. H., Kent R. D., Kent J. F., Milenkovic P., Workinger M. S. FORMOFFA: An automated formant, moment, fundamental frequency, amplitude analysis of normal and disordered speech // Clinical Linguistics & Phonetics. 1996. Vol. 10. Is. 1. P. 31–54. DOI: 10.3109/02699209608985160.

12. Von Neumann T., Boeddeker C., Kinoshita K., Delcroix M., Haeb-Umbach R. On word error rate definitions and their efficient computation for multi-speaker speech recognition systems // 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. DOI: 10.1109/icassp49357.2023.10094784.

13. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // Advances in Neural Information Processing Systems (NeurIPS). 2020. Vol. 33. P. 12449–12460. URL: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf (дата обращения: 26.03.2026).

14. Morris T. P., White I. R., Crowther M. J. Using simulation studies to evaluate statistical methods // Statistics in Medicine. 2019. Vol. 38. Is. 11. P. 2074–2102. DOI: 10.1002/sim.8086.

15. Kewley-Port D., Watson C. S. Formant-frequency discrimination for isolated English vowels // The Journal of the Acoustical Society of America. 1994. Vol. 95. Is. 1. P. 485–496. DOI: 10.1121/1.410024.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.