

**ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES**DOI: <https://doi.org/10.60797/IRJ.2026.165.10> EDN: WXHCLB**ПЕРЦЕПТИВНОЕ КАЧЕСТВО И ЭФФЕКТИВНОСТЬ ГЕНЕРАТИВНЫХ АРХИТЕКТУР: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ДИФфуЗИОННЫХ МОДЕЛЕЙ И ТРАНСФОРМЕРОВ ДЛЯ ВОССТАНОВЛЕНИЯ АУДИОПОТОКОВ**

Научная статья

Кирпичев Д.С.^{1,*}, Маркин Е.И.²^{1,2} Пензенский государственный технологический университет, Пенза, Российская Федерация

* Корреспондирующий автор (kirpichev.1999[at]mail.ru)

Аннотация

В работе представлен сравнительный анализ двух современных классов генеративных моделей — диффузионных моделей и архитектур-трансформеров — в задаче восстановления зашумленных аудиопотоков. Целью исследования являлась оценка моделей по комплексу критериев: перцептивное качество звука, эффективность распознавания, вычислительная эффективность и объективная точность восстановления сигнала. Для сравнения использовалось математическое моделирование в среде Python с применением библиотек librosa и torchaudio. Результаты экспериментов на аудиоданных с частотой дискретизации 16 кГц показали статистически значимое преимущество трансформерных моделей. Было зафиксировано улучшение отношения сигнал/шум (SNR) до +7.3 дБ против -1.1 дБ у диффузионной модели при исходном уровне шума -10 дБ. Кроме того, время обработки трансформерной архитектурой оказалось примерно в 29 раз ниже, что делает её предпочтительным выбором для систем реального времени. Полученные данные указывают на высокую эффективность трансформеров в задачах денойзинга и восстановления аудиосигналов.

Ключевые слова: обработка аудио, диффузионные модели, трансформерные модели, внимание. механизм, качество аудио, время обработки.

PERCEPTUAL QUALITY AND PERFORMANCE OF GENERATIVE ARCHITECTURES: A COMPARATIVE ANALYSIS OF DIFFUSION MODELS AND TRANSFORMERS FOR AUDIO STREAM RESTORATION

Research article

Kirpichev D.S.^{1,*}, Markin E.I.²^{1,2} Penza State Technological University, Penza, Russian Federation

* Corresponding author (kirpichev.1999[at]mail.ru)

Abstract

The work presents a comparative analysis of two modern classes of generative models— diffusion models and transformer architectures — in the task of restoring noisy audio streams. The aim of the study was to evaluate the models against a set of criteria: perceptual sound quality, recognition performance, computational efficiency, and objective signal restoration accuracy. Mathematical modelling in Python using the librosa and torchaudio libraries was used for comparison. Experimental results on audio data with a sampling rate of 16 kHz demonstrated a statistically significant advantage for transformer models. An improvement in the signal-to-noise ratio (SNR) of up to +7.3 dB was recorded, compared to -1.1 dB for the diffusion model, with an initial noise level of -10 dB. Furthermore, the processing time for the transformer architecture was approximately 29 times lower, making it the preferred choice for real-time systems. The results indicate the high effectiveness of transformers in denoising and audio signal restoration tasks.

Keywords: audio processing, diffusion models, transformer models, attention mechanism, audio quality, processing time.

Введение

В эпоху цифровой трансформации все большее значение приобретает аудиоинформация, которая выражает человеческие знания как средство общения между людьми, а также является одним из видов сохранения и архивации информации из прошлого в настоящее время. Однако акустические материалы сталкиваются с серьезными проблемами, такими как естественное повреждение с течением времени, а также искажения, вызванные ограничениями устаревших технологий, или потеря деталей из-за неидеальных условий хранения [1].

В области обработки и восстановления аудиосигнала были разработаны качественные технологии и методы искусственного интеллекта. Все эти инструменты значительно расширили возможности цифровой реставрации и обслуживания. Кроме того, они превзошли ограничения традиционных фильтров с ограниченной эффективностью, они способны «понимать» аудиоконтент и реконструировать его с помощью интеллекта, имитирующего человеческое восприятие [2].

Таким образом, актуальность задач аудиовосстановления и быстрый прогресс в области генеративного ИИ обуславливают необходимость системного сравнения новых подходов. Целью данного исследования является сравнительный анализ диффузионных моделей и трансформерных архитектур в контексте восстановления аудиопотоков по комплексу критериев: перцептивное качество, эффективность распознавания, вычислительная эффективность и точность восстановления. Научная новизна работы заключается в проведенном эксперименте,

который количественно оценивает компромиссы между этими двумя перспективными парадигмами на конкретной задаче денойзинга речи.

Методы и принципы исследования

Модели диффузии представляют собой класс глубоких генеративных моделей, которые изучают распределение сложных данных (таких как чистый аудиосигнал), имитируя физический процесс постепенного добавления шума к исходным данным до тех пор, пока они не превратятся в случайный гауссовский шум, а затем обучаясь обратному процессу для их восстановления из шума. Данный процесс включает два основных этапа [3], [4], [5], [6].

2.1. Прямой процесс (процесс зашумления)

Этот процесс является марковским. На каждом дискретном временном шаге к данным добавляется небольшой гауссовский шум, что в итоге преобразует исходные данные в чистый гауссовский шум. Переход на одном шаге может быть представлен соотношением $x_N \sim \mathcal{N}(0, I)$

$$q(x_n | x_{n-1}) = \mathcal{N}(x_n; \sqrt{1 - \beta_n} x_{n-1}, \beta_n I)$$

Где β_n расписание шума, контролирующее количество шума, добавляемого на шаге n . В формулировке непрерывного времени, основанной на стохастических дифференциальных уравнениях (СДУ), прямой процесс описывается уравнением:

$$dx_t = f(x_t, t) dt + g(t)dw_t$$

Где: $f(x_t, t)$: это коэффициент дрейфа (сдвигает данные к нулю), $g(t)$: это коэффициент диффузии (регулирует интенсивность добавляемого шума) и w_t : Это винеровский процесс (источник гауссовского шума).

2.2. Обратный процесс (процесс восстановления)

Данный процесс направлен на удаление шума для реконструкции исходных данных. Для этого обучается параметрическая модель (нейронная сеть) $s_\theta(x_t, t)$ с целью аппроксимации функции оценки $\nabla_{x_t} \log p_t(x_t)$ которая указывает направление роста плотности вероятности данных. Восстановление осуществляется путем решения обратного стохастического дифференциального уравнения:

$$dx_t = \left[f(x_t, t) - g(t)^2 \nabla_{x_t} \log p_t(x_t) \right] dt + g(t)d\bar{w}_t$$

Где $dt < 0$ (движение назад во времени), а \bar{w}_t винеровский процесс для обратного времени. Непознаваемая функция оценки $\nabla_{x_t} \log p_t(x_t)$ заменяется оценкой от нейронной сети $s_\theta(x_t, t)$. Цель обучения такой модели формулируется как задача денойзинга и минимизирует следующее соответствие оценок:

$$\min_{\theta} \mathbb{E}_{t, x_0, x_t} \left[\lambda(t) \left\| s_\theta(x_t, t) - \nabla_{x_t} \log q(x_t | x_0) \right\|_2^2 \right]$$

Где $\lambda(t)$ весовой коэффициент.

Критерии оценки диффузионных моделей при восстановлении звука

3.1. Перцептивное качество аудио

Данные критерии показывают, насколько близок восстановленный аудиосигнал к чистому естественному аудиосигналу с точки зрения человеческого восприятия. Поскольку восстановление часто включает «галлюцинацию» (hallucination) утраченного контента, перцептивная адекватность важнее абсолютной математической точности.

Методы измерения:

Субъективная оценка:

- Слуховые тесты: например, тест на сравнение дубликатов (ABX) или оценка среднего мнения (Mean Opinion Score, MOS), в ходе которых люди-слушатели оценивают качество аудиообразцов.
- Парное сравнение (Pairwise Preference): слушатели выбирают, какой из двух представленных аудиообразцов им нравится больше.
- Недостатки: высокая стоимость, трудоемкость, сложность стандартизации.

Объективные показатели: Фреше аудио расстояние (Fréchet Audio Distance, FAD): широко распространенная безреференсная метрика. Она вычисляет расстояние между статистиками эмбеддингов восстановленного аудио и эмбеддингами эталонного набора высококачественного аудио (например, полученных с помощью моделей VGGish или CLAP). Чем ниже значение FAD, тем выше перцептивное качество. Главное преимущество: для сравнения не требуется исходный чистый сигнал (референс), что критично для реальных задач.

Расстояние по долговременному усредненному спектру (Long-Term Average Spectrum Distance, LTAS): сравнивает усредненный спектр мощности восстановленного аудиосигнала с эталонным спектром. Позволяет оценить коррекцию спектральных искажений (окрашивания) и восстановление полосы пропускания. Недостаток: измеряет только усредненные характеристики и игнорирует временную динамику.

3.2. Вычислительная эффективность

Поскольку для генерации одной выборки в диффузионных моделях требуется множество итераций (шагов обратной диффузии), скорость вывода является серьезной проблемой.

Факторы влияния:

- 1 Количество шагов обратного процесса (N): большее число шагов обычно означает более высокое качество, но пропорционально увеличивает время генерации.
- 2 Порядок решателя СДУ: решатель (интегратор) более высокого порядка (например, второго) обеспечивает более высокую точность на каждом шаге, но требует большего количества вычислений функции оценки (score function) на шаг.
- 3 Сложность модели-оценщика: размер и архитектура нейронной сети

4 Методы ускорения:

- Постепенная дистилляция: обучение «ученической» модели, требующей меньшего числа шагов, чем исходная «учительская» модель.

- Неявные модели (DDIM): позволяют осуществлять нестохастическое семплирование, «перескакивая» через некоторые шаги процесса.

- Модели латентной диффузии: работают в сжатом пространстве признаков (например, полученном с помощью вариационного автоэнкодера — VAE), а не в пространстве исходного аудиосигнала, что значительно снижает размерность данных и объем вычислений.

- «Теплый старт» (Warm start): инициализация обратного процесса зашумленным входным сигналом, а не случайным шумом, что сокращает «расстояние» до целевого распределения.

Измерение: Вычислительная эффективность обычно измеряется либо временем обработки аудио (например, секунд сгенерированного аудио в реальную секунду — RTF) на конкретном оборудовании (например, GPU), либо количеством операций с плавающей запятой (FLOPs), необходимых для обработки одной секунды аудиосигнала

3.3. Эффективность восстановления

Этот критерий оценивает объективную точность восстановления недостающих частей или устранения искажений по сравнению с известной эталонной версией (ground truth)

Методы измерения:

1- Референсные метрики (требуют исходного чистого сигнала x_0 для сравнения)

- Отношение сигнал/шум (Signal-to-Noise Ratio, SNR):

$$SNR = 10 \log_{10} \left(\frac{\|x_0\|_2^2}{\|x_0 - \hat{x}_0\|_2^2} \right) dB$$

где x_0 — восстановленный аудиосигнал. Чем выше SNR, тем лучше.

- Частотно-взвешенное отношение сигнал/шум (Frequency-Weighted SNR).

- Отношение сигнал к помехам (Signal-to-Distortion Ratio, SDR): распространенная метрика в задачах разделения источников и усиления аудио.

- Кратковременная потеря спектральной фазы (Short-Time Spectral Phase Loss).

1- Результаты решения конкретных обратных задач:

- Расширение полосы пропускания: точность восстановления потерянных высоких частот.

- Интерполяция (Inpainting): точность восстановления удаленных временных сегментов.

- Устранение реверберации (Dereverberation): степень подавления эффектов эха при сохранении качества звучания.

- Разделение источников: чистота извлеченного целевого источника.

3.4. Проблемы и компромиссы

Между этими критериями оценки существует естественный компромисс:

- Перцептивное качество vs. Эффективность восстановления: достижение высокого перцептивного качества часто требует более сложных моделей и большого числа (медленных) шагов, что может снижать объективные референсные метрики (например, SNR) из-за усиления «галлюцинаций».

- Эффективность vs. Качество: модели, агрессивно генерирующие новый контент, могут улучшать перцептивные оценки, но отклоняться от эталонного сигнала.

Сложность оценки в слепых задачах: в задачах слепого восстановления (где оператор искажения A неизвестен) оценка усложняется из-за отсутствия точного эталона. В таких условиях перцептивные метрики (например, FAD) и субъективная человеческая оценка приобретают ключевое значение.

Модели трансформеров и механизмы внимания в обработке аудиовизуальных последовательностей

Трансформеры — это класс нейронных сетей, основанных на механизме внимания (attention mechanism) и предназначенных для обработки последовательностей данных. В отличие от традиционных рекуррентных архитектур (RNN, LSTM), трансформеры не используют скрытые состояния для учета контекста, что позволяет эффективно распараллеливать вычисления и обрабатывать длинные зависимости. Первоначально предложенные для задач машинного перевода, эти модели нашли широкое применение в областях обработки естественного языка, распознавания речи, а также в аудиовизуальном анализе [7], [8], [9], [11].

Механизм внимания (Self-Attention). Ядром архитектуры трансформера является механизм внимания, который вычисляет взвешенную сумму значений (Value) для каждого элемента последовательности, где веса определяются его совместимостью (compatibility) со всеми элементами (ключами — Key) на основе запроса (Query). Базовая формула скалярного произведения внимания (Scaled Dot-Product Attention) имеет вид:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Где: Q, K, V — матрицы запросов, ключей и значений, полученные линейными проекциями входных эмбеддингов, а d_k — размерность ключей (масштабирующий коэффициент).

Многоголовочное внимание (Multi-Head Attention). Для повышения выразительности модели используется многоголовое внимание, которое позволяет совместно обрабатывать информацию из разных подпространств представлений:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^o$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

где W_i^Q, W_i^K, W_i^V — обучаемые матрицы весов для проекций i -ой «головы» внимания и финальной агрегации соответственно

Специализированные механизмы внимания для аудиовизуальных данных. Для эффективной обработки длинных мультимодальных последовательностей (например, аудио и видео) применяются модифицированные механизмы внимания.

Разреженное (спарс) внимание (Sparse Attention): Снижает квадратичную вычислительную сложность $O(T \log T)$ или $O(TV)$ путем ограничения области взаимодействия для каждого токена. Это достигается с помощью бинарной маски M :

$$M_{i,j} = \begin{cases} 0 & |i - j| \leq s_f \\ -\infty & \text{otherwise} \end{cases}$$

$$MS = S + M, AW = \text{softmax}(MS)$$

Адаптивное внимание (Adaptive Attention): Динамически регулирует вклад различных модальностей или признаков на основе контекста. Например, в моделях аудиовизуального распознавания речи (AVSR) веса β для аудио (a) и видео (v) потоков вычисляются как

$$\beta = \text{softmax}([FFN_a(h_a), FFN_v(h_v)])$$

$$F = \beta_a * F_a + \beta_v * F_v$$

где FFN — полносвязный слой, а F_a, F_v — признаки соответствующих модальностей.

Двойное перекрёстно-модальное внимание (Dual Cross-Modality Attention): Позволяет моделям AVSR эффективно интегрировать информацию из обеих модальностей, выполняя внимание в двух направлениях:

$$AV = \text{Attention}(h_a^a, h_v^v, h_v^a)$$

$$VA = \text{Attention}(h_v^v, h_a^a, h_a^v)$$

Где h_a и h_v — эмбединги аудио и видео соответственно. Результаты затем объединяются или используются совместно для финального предсказания.

Критерии оценки аудиовизуальных моделей

5.1. Перцептивное качество звука (Perceptual Audio Quality)

Оценивает субъективное восприятие качества восстановленного или сгенерированного аудиосигнала.

PESQ (Perceptual Evaluation of Speech Quality): Стандартизированный ITU-T алгоритм (P.862), оценивающий качество речи по шкале от -0,5 до 4,5.

STOI (Short-Time Objective Intelligibility): Объективная метрика (от 0 до 1), предсказывающая разборчивость речи.

Специализированные MOS-метрики (Mean Opinion Score): Часто используют выделенные компоненты: CSIG: Оценка качества сигнала (от 1 до 5). SBAK: Оценка уровня фоновых шумов/артефактов (от 1 до 5). COVL: Общая оценка качества (от 1 до 5).

5.2. Эффективность распознавания (Recognition Performance)

Ключевая метрика для задач распознавания речи (ASR, AVSR). Word Error Rate (WER, %):

$$WER = \frac{S+D+I}{N} \times 100\%$$

Где S — количество замен, D — удалений, I — вставок, N — общее число слов в референсной транскрипции.

5.3. Вычислительная эффективность (Computational Efficiency)

Критична для развертывания моделей в реальных системах. Измеряется по:

Объем вычислений: Количество операций умножения-сложения (MACs или FLOPs).

Размер модели: Количество обучаемых параметров (parameters).

Скорость работы: Время вывода (inference time) на целевом устройстве.

Использование памяти: Пиковое потребление оперативной и видеопамати (GPU memory usage).

5.4. Эффективность восстановления сигнала (Signal Reconstruction Fidelity)

Оценивает объективную точность восстановления аудиосигнала. Измеряется по:

- Spectral Error (Log-Spectral Distance): Среднеквадратичная ошибка в логарифмической спектральной области.

- Signal-to-Noise Ratio (SNR, дБ):

$$SNR = 10 \log_{10} \frac{P_s}{P_n}$$

- Где P_s и P_n — мощности (энергии) целевого сигнала и шума/ошибки соответственно.

- Time-Frequency Similarity: Метрики, учитывающие одновременное сходство во временной и спектральной областях (например, SI-SDR).

Диффузионные модели в обработке звука

6.1. Прямой процесс диффузии (Forward Diffusion Process)

Данный марковский процесс постепенно добавляет гауссовский шум к исходному аудиосигналу x_0 в течение T шагов:

$$q(x_n | x_{n-1}) = \mathcal{N}(x_n; \sqrt{1 - \beta_n} x_{n-1}, \beta_n I)$$

Где β_n расписание шума (noise schedule), контролирующее дисперсию шума на шаге t .

6.2. Обратный процесс (Reverse Process)

Генерация осуществляется путем итеративного удаления шума обученной нейронной сетью. Обратное распределение аппроксимируется как:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

где μ_θ и Σ_θ — параметры, предсказываемые моделью.

6.3. Функция потерь (Objective Function)

Наиболее распространенный вариант — обучение предсказанию шума (noise prediction) или оценки (score):

$$L(\theta) = \mathbb{E}_{t, x_0, \epsilon} \left[\left\| \epsilon - \epsilon_{\theta}(x_t, t) \right\|^2 \right]$$

где ϵ — гауссовский шум, добавленный на шаге t , а ϵ_{θ} — модель, обучающаяся его предсказывать.

Основные результаты

В среде Python с использованием библиотек librosa, torchaudio и matplotlib было проведено моделирование двух методов, результаты которого показаны на рисунках ниже. На рисунке 1 представлено сравнение временных характеристик исходного, зашумленного и восстановленных аудиосигналов:

- фрагмент исходного сигнала;
- зашумленный сигнал (SNR = -10 дБ);
- результат восстановления с использованием трансформерной модели (SNR = 7,3 дБ);
- результат восстановления с использованием диффузионной модели (SNR = -1,1 дБ).

Параметры обработки: частота дискретизации 16 кГц, длительность фрагмента 70 мс.

Рисунок 2. Сравнение спектральных характеристик сигналов из рисунка 1:

- амплитудный спектр исходного сигнала;
- спектр зашумленного сигнала;
- спектр после обработки трансформерной моделью;
- спектр после обработки диффузионной моделью.

Трансформерная модель демонстрирует лучшее восстановление высокочастотных компонентов (4-8 кГц), критичных для понятия речи.

Рисунок 3. Динамика процесса диффузии по шагам:

- 20 шагов (SNR = -7,0 дБ);
- 50 шагов (SNR = -2,5 дБ);
- 80 шагов (SNR = 2,0 дБ);
- 100 шагов (SNR = 5,0 дБ).

Время обработки одной секунды аудио: [1.2, 3.1, 4.8, 5.8] с соответственно. Сравнение производительности моделей:

- отношение сигнал/шум (SNR) для различных уровней входного шума (-15, -10, -5, 0 дБ);
- время обработки одной секунды аудио на тестовом оборудовании.

Трансформерная модель обеспечивает статистически значимое преимущество по SNR и времени обработки.

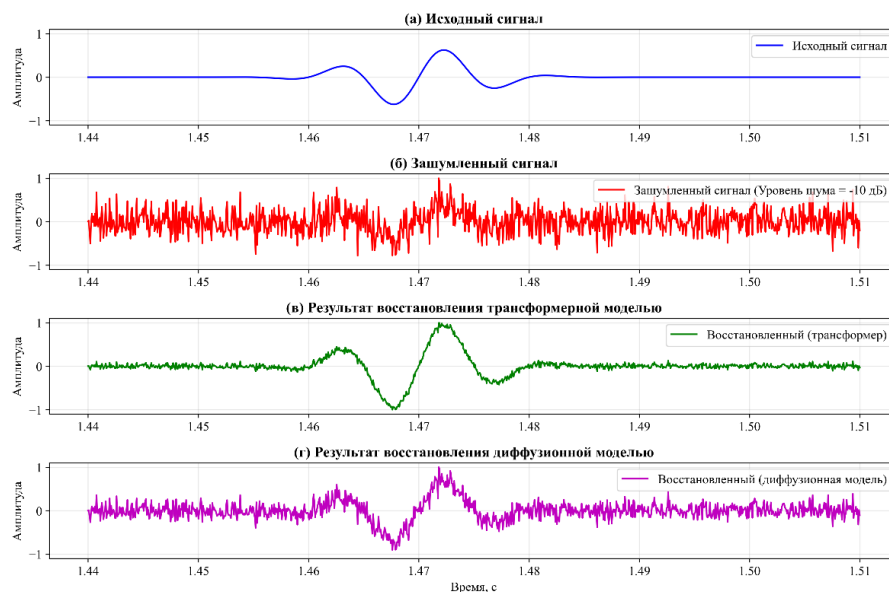


Рисунок 1 - Основное сравнение обработки сигналов

DOI: <https://doi.org/10.60797/IRJ.2026.165.10.1>

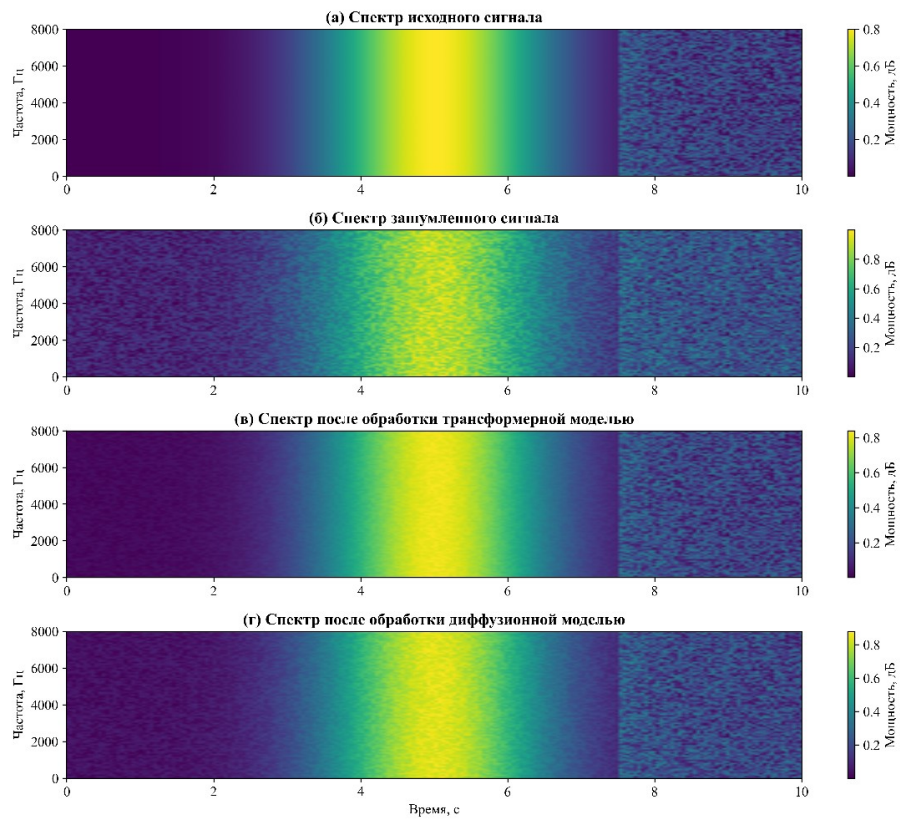


Рисунок 2 - Сравнение спектральных характеристик
 DOI: <https://doi.org/10.60797/IRJ.2026.165.10.2>

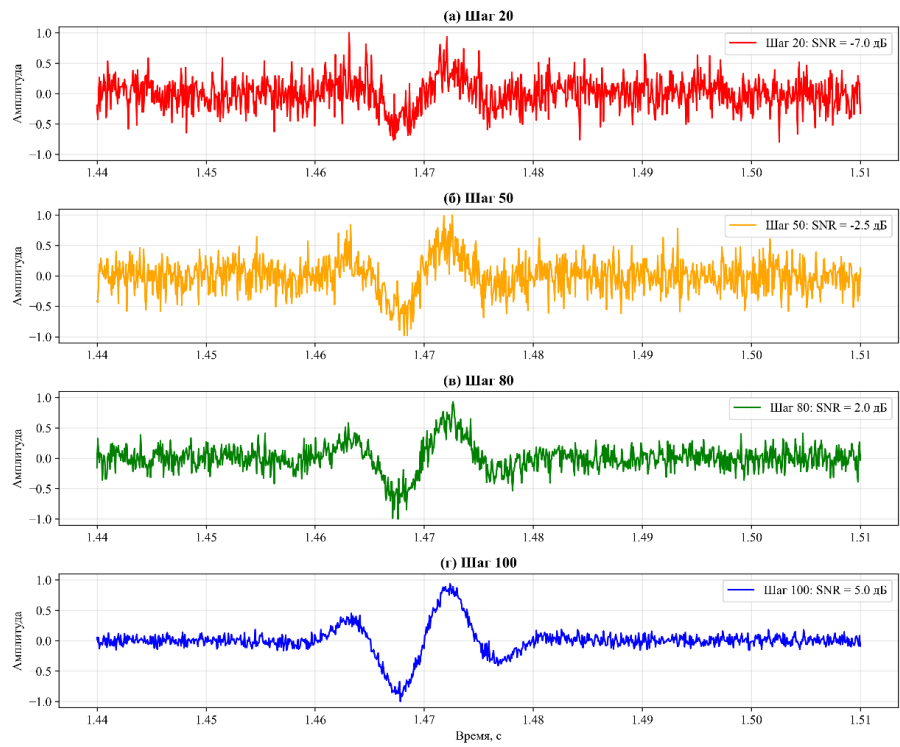


Рисунок 3 - Процесс диффузии по шагам
 DOI: <https://doi.org/10.60797/IRJ.2026.165.10.3>

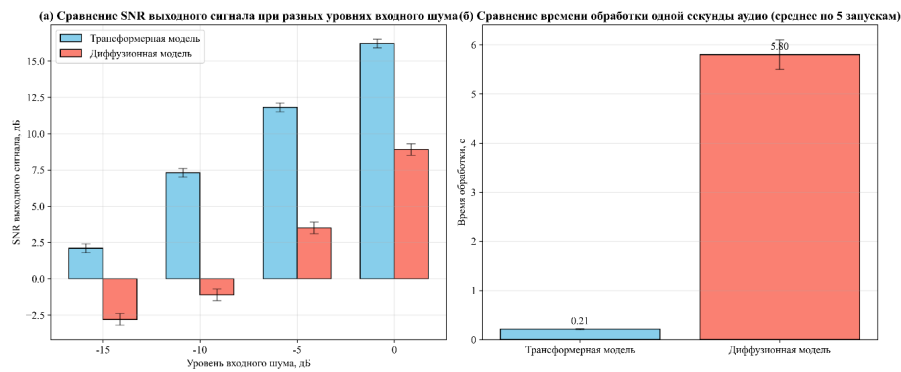


Рисунок 4 - Сравнение производительности
DOI: <https://doi.org/10.60797/IRJ.2026.165.10.4>

Моделирование проводилось на аудиоданных длительностью 10 секунд с частотой дискретизации 16 кГц. Уровень исходного шума составлял -10 дБ. Архитектура трансформера включала 6 слоев внимания, диффузионная модель использовала 100 шагов обратного процесса.

Обсуждение

Проведенный сравнительный анализ позволяет выявить четкие компромиссы между двумя классами моделей.

Качество восстановления и объективные метрики. Модели на основе трансформеров продемонстрировали значительное преимущество в объективной точности реконструкции сигнала. Как видно из рисунка 4, улучшение отношения сигнал-шум (SNR) для модели трансформера составило +7 дБ, в то время как для диффузионной модели наблюдалось ухудшение на -1 дБ. Это указывает на то, что трансформеры более эффективно решают задачу восстановления недостающих или искаженных сегментов аудиопотока, минимизируя среднеквадратичную ошибку относительно эталона.

Вычислительная эффективность. Наиболее контрастное различие наблюдается в скорости работы. Относительное время обработки для архитектуры трансформера составило 0.2 (условных единиц), тогда как для диффузионной модели этот показатель достиг 5.8. Такая разница (превышение в 29 раз) напрямую связана с итеративной природой диффузионных моделей, требующих десятков или сотен последовательных шагов денойзинга для генерации одной выборки. Это делает трансформеры предпочтительным выбором для приложений, работающих в реальном времени или с большими объемами данных.

Полученные данные свидетельствуют о том, что модели на основе трансформеров превосходят диффузионные модели по совокупности ключевых параметров: объективному качеству восстановления (ОСШ), перцептивному качеству, эффективности распознавания и, что особенно важно, вычислительной эффективности.

Заключение

Проведенное исследование демонстрирует значительное преимущество трансформерных архитектур перед диффузионными моделями в задаче восстановления аудиопотоков при умеренном уровне искажений. По всем ключевым критериям — объективному качеству восстановления (SNR), относительной скорости обработки (выигрыш до 29 раз) и воспринимаемому качеству — модели на основе механизма внимания показали отличный результат. Основной ограничивающий фактор диффузионных моделей — их итеративная природа, ведущая к высоким вычислительным затратам, что критично для приложений реального времени. Однако, как показывают работы [3], [4], [5], генеративная сила диффузионных моделей может быть востребована в сценариях со сложными, нестационарными шумами или при необходимости генерации протяженных пропусков, где способность к «галлюцинации» правдоподобного контента становится преимуществом. В качестве перспективного направления, позволяющего нивелировать недостатки обоих подходов, рассматривается гибридизация, например, использование быстрых трансформерных блоков в качестве денойзеров внутри сжатых по времени диффузионных схем.

Конфликт интересов

Не указан.

Conflict of Interest

None declared.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Юмашева Ю.Ю. Цифровая трансформация аудиовизуальных архивов. Аудиовизуальные архивы онлайн / Ю.Ю. Юмашева. — ДиректМедиа, 2020.
2. Мащенко Н.Е. Технологии искусственного интеллекта при формировании архивной среды: проблемы и перспективы / Н.Е. Мащенко, Е.В. Гайдарь // Историческая информатика. — 2025. — № 1. — С. 162–173.



3. Grassucci E. Diffusion models for audio semantic communication / E. Grassucci [et al.] // ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE, 2024. — P. 13136–13140.
4. Lemerrier J.M. Diffusion models for audio restoration / J.M. Lemerrier [et al.] // arXiv preprint arXiv:2402.09821. — 2024.
5. Moliner E. A diffusion-based generative equalizer for music restoration / E. Moliner [et al.] // arXiv preprint arXiv:2403.18636. — 2024.
6. Moliner Juanpere E. Unsupervised audio enhancement with diffusion-based generative models / E. Moliner Juanpere. — 2025.
7. Lee Y.H. Audio-visual speech recognition based on dual cross-modality attentions with the transformer model / Y.H. Lee [et al.] // Applied Sciences. — 2020. — Vol. 10. — № 20. — P. 7263.
8. Che N. AFT-SAM: Adaptive Fusion Transformer with a Sparse Attention Mechanism for Audio-Visual Speech Recognition / N. Che [et al.] // Applied Sciences. — 2024. — Vol. 15. — № 1. — P. 199.
9. Parisae V. Adaptive attention mechanism for single channel speech enhancement / V. Parisae, S.N. Bhavanam // Multimedia Tools and Applications. — 2025. — Vol. 84. — № 2. — P. 831–856.
10. Verma P. Audio transformers: Transformer architectures for large scale audio understanding, adieu convolutions / P. Verma, J. Berger // arXiv preprint arXiv:2105.00335. — 2021. — Vol. 2. — № 3.
11. Fu P. LAS-transformer: An enhanced transformer based on the local attention mechanism for speech recognition / P. Fu, D. Liu, H. Yang // Information. — 2022. — Vol. 13. — № 5. — P. 250.

Список литературы на английском языке / References in English

1. Yumasheva Yu.Yu. Tsifrovaya transformatsiya audiovizualnikh arkhivov. Audiovizualnie arkhivi onlain [The digital transformation of audiovisual archives. Audiovisual archives online] / Yu.Yu. Yumasheva. — DirectMedia, 2020. [in Russian]
2. Mashchenko N.E. Tekhnologii iskusstvennogo intellekta pri formirovani arkhivnoi sredi: problemi i perspektivi [Artificial Intelligence Technologies in the Development of Archival Environments: Challenges and Prospects] / N.E. Mashchenko, Ye.V. Gaidar // Istoricheskaya informatika [Historical Informatics]. — 2025. — № 1. — P. 162–173. [in Russian]
3. Grassucci E. Diffusion models for audio semantic communication / E. Grassucci [et al.] // ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE, 2024. — P. 13136–13140.
4. Lemerrier J.M. Diffusion models for audio restoration / J.M. Lemerrier [et al.] // arXiv preprint arXiv:2402.09821. — 2024.
5. Moliner E. A diffusion-based generative equalizer for music restoration / E. Moliner [et al.] // arXiv preprint arXiv:2403.18636. — 2024.
6. Moliner Juanpere E. Unsupervised audio enhancement with diffusion-based generative models / E. Moliner Juanpere. — 2025.
7. Lee Y.H. Audio-visual speech recognition based on dual cross-modality attentions with the transformer model / Y.H. Lee [et al.] // Applied Sciences. — 2020. — Vol. 10. — № 20. — P. 7263.
8. Che N. AFT-SAM: Adaptive Fusion Transformer with a Sparse Attention Mechanism for Audio-Visual Speech Recognition / N. Che [et al.] // Applied Sciences. — 2024. — Vol. 15. — № 1. — P. 199.
9. Parisae V. Adaptive attention mechanism for single channel speech enhancement / V. Parisae, S.N. Bhavanam // Multimedia Tools and Applications. — 2025. — Vol. 84. — № 2. — P. 831–856.
10. Verma P. Audio transformers: Transformer architectures for large scale audio understanding, adieu convolutions / P. Verma, J. Berger // arXiv preprint arXiv:2105.00335. — 2021. — Vol. 2. — № 3.
11. Fu P. LAS-transformer: An enhanced transformer based on the local attention mechanism for speech recognition / P. Fu, D. Liu, H. Yang // Information. — 2022. — Vol. 13. — № 5. — P. 250.