НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2026 Том 26 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2026 Vol. 26 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

ОБЗОРНЫЕ СТАТЬИ

REVIEW PAPERS

doi: 10.17586/2226-1494-2026-26-4-673-682
УДК 004.5, 004.934

Нейросетевые методы улучшения качества речевых сигналов: архитектурные парадигмы и тенденции развития

Денис Викторович Иванько✉

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

ivanko.d@iias.spb.su✉, https://orcid.org/0000-0003-0412-7765

Аннотация

Введение. Представлен систематический обзор современных нейросетевых методов улучшения качества речевых сигналов, предназначенных для повышения разборчивости и качества речи в условиях акустических искажений. Рассмотренные методы могут найти применение в системах голосового управления, телекоммуникациях, слуховых аппаратах и интерфейсах человеко-машинного взаимодействия. **Методы.** Рассмотрены ключевые архитектурные подходы, включая классические рекуррентные и сверточные сети, а также современные гибридные архитектуры с механизмами внимания (Transformer, Conformer), модели состояний (Mamba) и усовершенствованные рекуррентные блоки (xLSTM). **Основные результаты.** Показаны достоинства и недостатки различных архитектур с точки зрения качества восстановления речи и вычислительной эффективности. Выделены конкретные проблемы существующих методов, включая высокую ресурсоемкость и недостаточную обобщающую способность в нестационарных шумовых условиях. **Обсуждение.** Показана необходимость дальнейших исследований в области создания легковесных моделей для мобильных устройств, разработки методов многоцелевого подавления искажений, а также интеграции нейросетевого подавления шума с генеративными моделями для достижения нового качества восстановления речевых сигналов.

Ключевые слова

улучшение качества речевых сигналов, подавление шума, глубокое обучение, трансформеры, Mamba, xLSTM, нейросетевые архитектуры, обзор

Благодарности

Работы в разделе «Новая парадигма оценки SE-систем» выполнены в рамках бюджетной темы № FFZF-2025-0003. Все остальные разделы выполнены в рамках поддержки гранта РНФ 25-71-00093.

Ссылка для цитирования: Иванько Д.В. Нейросетевые методы улучшения качества речевых сигналов: архитектурные парадигмы и тенденции развития // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 4. С. 673–682. doi: 10.17586/2226-1494-2026-26-4-673-682

Deep learning for speech enhancement: architectures, paradigms, and emerging trends

Denis V. Ivanko✉

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation

ivanko.d@iias.spb.su✉, https://orcid.org/0000-0003-0412-7765

Abstract

A systematic review of modern neural network methods for Speech Enhancement is presented, aimed at improving speech intelligibility and quality under acoustic distortions. The reviewed methods can be applied in voice control systems, telecommunications, hearing aids, and human–machine interaction interfaces. Key architectural approaches are considered, including classical recurrent and convolutional networks as well as modern hybrid architectures with

attention mechanisms (Transformer, Conformer), state-space models (Mamba), and advanced recurrent blocks (xLSTM). The advantages and disadvantages of different architectures are shown in terms of speech restoration quality and computational efficiency. Specific problems of existing methods are highlighted, including high computational cost and insufficient generalization capability under non-stationary noise conditions. The need for further research in the development of lightweight models for mobile devices, multi-distortion suppression methods, and the integration of neural network noise suppression with generative models to achieve a new level of speech signal restoration quality is demonstrated.

Keywords
speech enhancement, noise suppression, deep learning, neural network architectures, transformers, Mamba, xLSTM, survey

Acknowledgements
The work in the "New Paradigm for Assessing SE Systems" section was completed under budget topic No. FFZF-2025-0003. All other sections were supported by RSF grant No. 25-71-00093.

For citation: Ivanko D.V. Deep learning for speech enhancement: architectures, paradigms, and emerging trends. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 4, pp. 673–682 (in Russian). doi: 10.17586/2226-1494-2026-26-4-673-682

Введение

Речевой сигнал в реальных условиях эксплуатации практически всегда подвержен воздействию акустических помех различной природы: стационарные и нестационарные шумы, реверберация, эхо, нелинейные искажения. Это приводит к снижению разборчивости речи, ухудшению восприятия и, как следствие, к падению эффективности работы систем голосовой связи, автоматического распознавания речи, верификации диктора и голосовых интерфейсов человеко-машинного взаимодействия. Улучшение качества речевых сигналов (Speech Enhancement, SE) является одной из фундаментальных задач цифровой обработки сигналов, имеющей критическое значение для широкого спектра приложений. За последнее десятилетие глубокое обучение произвело революцию в этой области, позволив достичь беспрецедентного качества подавления шумов в стационарных условиях [1–3].

Задача подавления шума имеет многолетнюю историю. Традиционные подходы, такие как спектральное вычитание [4], винеровская фильтрация [5] и субпространственные методы [6], демонстрируют ограниченную эффективность в условиях нестационарных шумов и при низких отношениях сигнал/шум. Значительный прорыв произошел с внедрением методов глубокого обучения, которые позволили перейти от классических алгоритмов к моделям, обучаемым на больших массивах данных [7, 8].

Однако, несмотря на впечатляющий прогресс, современные SE-системы сталкиваются с фундаментальными ограничениями. Традиционные подходы, демонстрирующие высокую эффективность на тестовых данных из того же распределения, что и обучающая выборка, часто показывают сильное падение качества при работе в реальных акустических сценариях [9–11].

Настоящая работа представляет систематический обзор современных нейросетевых методов SE. Цель работы — классифицировать существующие подходы, выявить ключевые тренды, проанализировать компромиссы между качеством восстановления и вычислительной эффективностью, а также определить перспективные направления дальнейших исследований.

Эволюция архитектурных подходов SE: от классических рекуррентных сетей к гибридным архитектурам

Эволюция нейросетевых методов SE насчитывает более десятилетия активных исследований. Общая эволюция архитектурных подходов к SE представлена на рисунке. За этот период сменилось несколько архитектурных парадигм — от полносвязных и сверточных нейронных сетей (Convolutional Neural Network, CNN) до рекуррентных архитектур (Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM)), трансформеров (Transformer, Conformer) и новейших моделей состояний (State Space Models (SSM), Mamba). Каждый следующий этап был обусловлен стремлением преодолеть ограничения предшественников, такие как недостаточная моделирующая способность, высокая вычислительная сложность или слабое обобщение модели на нестационарные акустические условия. Приведем результат систематического анализа ключевых архитектурных семейств, их достоинств, фундаментальных ограничений (рисунок).

Сверточные архитектуры. Данные архитектуры стали первым успешным примером глубокого обучения в задачах SE, продемонстрировав существенное преимущество перед классическими методами спектрального вычитания и адаптивной фильтрации. Работы [12, 13] показали, что полносвязные сети уступают сверточным благодаря присущей последним способности эффективно моделировать локальные частотно-временные зависимости в спектрограммах или непосредственно во временной области [14].

Архитектуры типа Wave-U-Net [15, 16] и Conv-TasNet [17, 18] достигли значительных успехов в обработке сигналов во временной области, обеспечивая конкурентоспособное качество разделения источников. Conv-TasNet, в частности, ввела концепцию сверточных кодировщиков-декодировщиков с разделением по полосам, что впоследствии было адаптировано для задач SE.

Вместе с тем сверточные архитектуры обладают фундаментальным ограничением: их поле восприятия расширяется линейно с глубиной сети даже при использовании расширенных сверток. Кроме того, как отмечается в [19], эффективность чистых CNN-архитектур

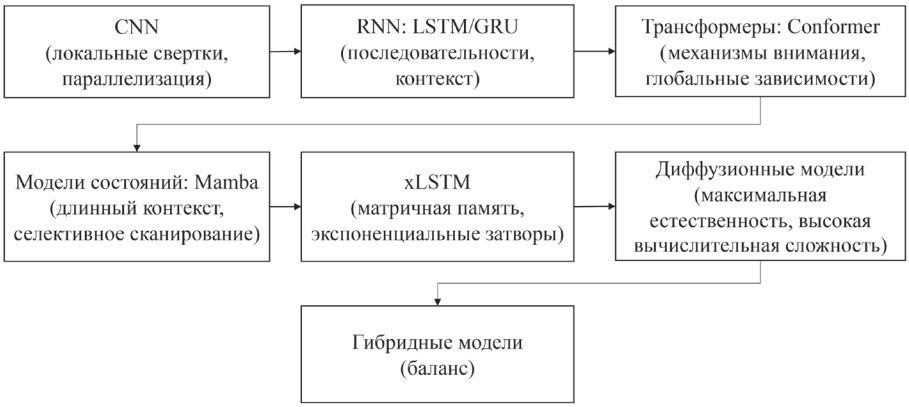

Рисунок. Эволюция архитектурных подходов к улучшению качества речевых сигналов. GRU — Gated Recurrent Unit; xLSTM — extended Long Short-Term Memory

Figure. Evolution of architectural approaches to speech enhancement

ограничена при работе с широкополосными сигналами (шире 16 кГц), где требуется более детальное частотное разрешение.

Рекуррентные сети. Отметим, что LSTM [20] на протяжении многих лет оставались доминирующим подходом для моделирования временных зависимостей в SE. Их принципиальное преимущество заключается в естественной обработке последовательностей переменной длины и линейной вычислительной сложности относительно числа временных шагов [21, 22].

Однако классические LSTM имеют ряд ограничений. Во-первых, они используют скалярное состояние ячейки, что ограничивает емкость долговременной памяти. Во-вторых, механизм затворов не позволяет пересматривать ранее принятые решения о сохранении или забывании информации [23]. В-третьих, последовательная природа рекуррентных сетей ограничивает степень параллелизации при обучении, что особенно критично при работе с длинными аудиопоследовательностями [24].

В последнее время наблюдается возрождение интереса к рекуррентным подходам, связанное с появлением архитектуры Extended LSTM (xLSTM) [25].

Экспериментальные результаты, представленные в работах [26, 27], показывают, что xLSTM способна достигать качества, сопоставимого с трансформерами, при сохранении линейной вычислительной сложности. В контексте SE это открывает перспективы создания моделей, которые учитывают длительный контекст (до 10–15 с речи) без квадратичного роста затрат.

Таким образом, эволюция рекуррентных сетей в SE движется от базовых LSTM к более мощным вариантам с расширенной памятью, что позволяет им конкурировать с более современными архитектурами.

Архитектуры на основе механизма внимания (Transformer, Conformer). Механизм самовнимания, лежащий в основе архитектуры Transformer [28], произвел революцию в обработке последовательностей, заменив рекуррентные связи прямыми взаимодействиями между всеми элементами последовательности. В контексте SE это означает, что модель может учитывать глобальные зависимости между произвольными временными или частотными позициями, что особенно важно для учета долгосрочного контекста на уровне всего высказывания.

Применение трансформеров в SE началось с работ [29], которые показали, что замена LSTM-слоев на блоки самовнимания приводит к улучшению качества восстановления речи, особенно в условиях сложных нестационарных шумов. Однако чистые трансформеры имеют один фундаментальный недостаток, а именно квадратичную вычислительную сложность относительно длины последовательности. Для типичных аудиозаписей длительностью 5–10 с при частоте дискретизации 16 кГц приводит к миллионам попарных взаимодействий, что делает обучение ресурсоемким [30].

Компромиссным решением стала гибридная архитектура Conformer (Convolution-augmented Transformer) [31, 32]. Conformer объединяет CNN для эффективного моделирования локальных паттернов (на масштабах 10–30 мс) и механизмы самовнимания для глобальных зависимостей (на масштабах всего высказывания). Такая комбинация позволяет сохранить преимущества обоих подходов, частично снижая вычислительную нагрузку за счет того, что CNN уменьшают пространственную размерность перед блоками внимания.

На бенчмарке VoiceBank+DEMAND Conformer демонстрирует одни из лучших результатов по метрикам Perceptual Evaluation of Speech Quality и Short-Time Objective Intelligibility [33, 34], превосходя как чистые CNN, так и LSTM-архитектуры. Тем не менее, проблема квадратичной сложности сохраняется для длинных записей, что стимулирует поиск альтернативных подходов с субквадратичной сложностью.

Модели состояний (SSM, Mamba). Модели состояний SSMпредставляют собой новый класс архитектур, которые объединяют преимущества CNN, RNN и трансформеров, избегая их ключевых недостатков. Особого внимания заслуживает архитектура Mamba [35], которая вводит механизм селективного сканирования. В отличие от классических SSM с фиксированными параметрами, Mamba адаптивно фильтрует информацию в зависимости от входных данных, что позволяет

модели динамически решать, какую часть контекста сохранять, а какую отбрасывать.

Первые результаты применения Mamba в задачах SE [36] подтвердили высокий потенциал этого подхода. В работах [37, 38] показано, что Mamba достигает результатов, сопоставимых с Conformer, при значительно меньшем количестве параметров (в 2–4 раза) и более высокой скорости инференса (в 3–5 раз для длинных записей). Особенно перспективным является использование двунаправленных конфигураций Mamba (Bi-Mamba) [39]. Таким образом, модели состояний, и Mamba в частности, рассматриваются в современных научных работах как один из наиболее многообещающих путей создания эффективных SE-систем с длинным контекстом, пригодных для работы на мобильных устройствах и в реальном времени.

От дискриминативных к генеративным подходам

Доминирующая на протяжении многих лет парадигма SE основана на дискриминативном (предиктивном) подходе. В рамках этой парадигмы модель обучается регрессии от зашумленного сигнала к чистому, минимизируя среднеквадратичную ошибку (Mean Squared Error, MSE), L1-расстояние или их подобные варианты [40]. Такой подход эффективен и вычислительно прост, однако имеет фундаментальное ограничение: минимизация MSE приводит к усреднению по всем возможным правдоподобным чистым сигналам. В результате восстановленная речь часто звучит сглаженной, теряет естественные микроколебания и содержит характерные артефакты [41].

Генеративные модели предлагают принципиально иную парадигму. Вместо оценки единственного «наиболее вероятного» чистого сигнала они моделируют условное распределение. Это позволяет генерировать различные правдоподобные варианты восстановления, что потенциально снижает артефакты усреднения и повышает естественность звучания [2].

Среди генеративных подходов наибольшее распространение получили два семейства: генеративно-состязательные сети (Generative Adversarial Networks, GANs) [42] и диффузионные вероятностные модели [43].

Speech Enhancement Generative Adversarial Network [8] была одной из первых работ, применивших GAN для улучшения речи. Генератор обучается восстанавливать чистый сигнал, а дискриминатор — отличать восстановленную речь от реальной чистой. Такой состязательный процесс стимулирует генератор производить статистически неотличимые от естественной речи образцы. Недостатком GAN является сложность обучения (нестабильность, режим коллапса) и возможность появления новых типов артефактов.

Диффузионные модели. В последние два года именно диффузионные модели привлекли наибольшее внимание исследователей [44, 45]. Они работают в два этапа: прямой процесс постепенно добавляет шум к чистому сигналу, а обратный процесс восстанавливает его, используя нейронную сеть для предсказания шума. Первые применения диффузионных моделей в SE [46] показали впечатляющую способность работать с различными типами искажений: аддитивным шумом, реверберацией, ограничением полосы и даже клиппированием.

Однако, как показали результаты соревнований по улучшению качества речевых сигналов URGENT Challenge 2025, диффузионные модели имеют три критических недостатка.

1) Языковая зависимость. Чисто генеративные SE-модели продемонстрировали существенные различия в качестве восстановления для разных языков (например, английского, китайского, немецкого), тогда как дискриминативные модели были значительно более стабильны.
2) Генеративные артефакты. Систематическое исследование, проведенное в [45], выявило специфические артефакты, характерные для диффузионных моделей: фонетические ошибки (замена одних фонем на другие), появление «музыкальных тонов» в паузах и нестабильность вокализованных сегментов.
3) Вычислительная сложность. Диффузионные модели требуют от 20 до 100 итераций обратного процесса для получения качественного результата. Это делает их непригодными для применения в реальном времени на мобильных устройствах и умных наушниках, где допустимая задержка составляет десятки миллисекунд.

Гибридные подходы: объединение дискриминативных и генеративных компонентов

Современный этап развития SE-архитектур характеризуется отказом от противопоставления дискриминативных и генеративных подходов в пользу их гибридизации. Общая идея заключается в том, чтобы использовать дискриминативную модель для быстрого получения начальной (грубой) оценки чистого сигнала, а затем уточнять ее генеративной сетью, которая исправляет артефакты и добавляет естественные детали.

Особого внимания заслуживает архитектура Dual-Path RNN (DPRNN) [47], которая объединяет внутрисегментную и межсегментную обработки. DPRNN разбивает длинную последовательность на короткие сегменты, обрабатывает каждый сегмент локальной RNN, а затем выполняет взаимодействие между сегментами глобальной RNN. Такой подход позволяет моделировать как локальные, так и глобальные зависимости при линейной сложности.

Ряд исследований [47] показал, что добавление механизмов самовнимания к RNN существенно улучшает кросс-корпусную генерализацию. Модели, обученные на одном корпусе (например, DNS Challenge), демонстрируют значительно меньшее падение качества при тестировании на другом (например, VoiceBank + DEMAND) благодаря способности внимания выделять инвариантные признаки. По итогам URGENT Challenge 2024–2026 годов именно гибридные решения заняли лидирующие позиции.

Сравнительный анализ всех рассмотренных архитектурных подходов в SE, представлен в таблице.

Таблица. Сравнительный анализ архитектурных подходов в SE
Table. Comparative analysis of architectural approaches in SE

Архитектура	Ключевая особенность	Преимущества	Ограничения	Качество восстановления	Пригодность для работы в реальном времени
CNN	Локальные свертки, иерархическое извлечение признаков	Эффективная параллелизация, инвариантность к сдвигам	Ограниченное рецептивное поле, слабое моделирование долгосрочного контекста	Среднее	Да
RNN (LSTM, GRU)	Рекуррентная обработка, затворы забывания/входа	Естественная работа с последовательностями переменной длины, линейная сложность	Скалярное состояние ячейки, невозможность пересмотра решений, слабая параллелизация	Высокое	Да
xLSTM	Экспоненциальные затворы, матричная память	Преодоление ограничений LSTM, качество на уровне трансформеров	Относительная новизна, меньшее количество бенчмарков	Очень высокое	Ограниченно
Transformer	Самовнимание, позиционное кодирование	Прямое моделирование глобальных зависимостей, параллелизация	Квадратичная сложность по длине последовательности, требовательность к памяти	Высокое	Да
Conformer	Свертка + самовнимание (гибрид CNN+Transformer)	локальные + глобальные зависимости (преимущество подходов CNN+Transformer)	Частичное сохранение квадратичной сложности, сложность обучения	Наивысшее	Да
Mamba (SSM)	Селективное сканирование, линейные модели состояний	Линейная сложность с длинным контекстом, быстрый инференс	Относительная новизна, требует адаптации под SE-задачи	Очень высокое	Да
GAN (генеративный)	Состязательное обучение генератора и дискриминатора	Естественное звучание, отсутствие сглаживания	Нестабильность обучения, артефакты	Естественное звучание	Ограниченно
Диффузионные вероятностные модели	Прямой/обратный диффузионный процесс	Наивысшая естественность, работа с сильными искажениями	Очень высокая вычислительная сложность (десятки итераций)	Наивысшее (естественность)	Нет
Гибридные	Дискриминативная начальная оценка + генеративное уточнение	Баланс качества и скорости, гибкость настройки	Сложность проектирования, два этапа обработки	Наивысшее	Частично

Новая парадигма оценки SE-систем

Долгое время прогресс в области улучшения качества речи оценивался в рамках узкоспециализированных задач. Типичный исследовательский сценарий выглядел следующим образом: модель обучается на одном-двух публичных корпусах (например, VoiceBank+DEMAND), а тестирование проводится на синтетических данных, сгенерированных по схожим принципам. Такой подход, несомненно, способствовал быстрому развитию методов, но породил системную проблему: модели, демонстрирующие впечатляющие результаты в лабораторных условиях, часто оказывались неработоспособными при столкновении с реальными акустическими сценариями.

В ответ на эти вызовы международным научным сообществом была инициирована серия соревнований URGENT (Universality, Robustness, and Generalizability of speech EnhancemeNT). Основная идея URGENT заключается в переходе от оценки моделей в узких, контролируемых условиях к комплексному тестированию универсальности, устойчивости и способности к генерализации на ранее не встречавшиеся типы искажений, языки и акустические сценарии.

Традиционные SE-бенчмарки, такие как DNS Challenge [48], CHiME [49], фокусировались на узких задачах: подавление аддитивного шума, подавление реверберации или их комбинация. URGENT Challenge предложил принципиально новую концепцию универсального SE, требующую от моделей способности работать с множеством типов искажений в рамках единой архитектуры.

Ключевое нововведение заключается в том, что одна и та же модель должна эффективно справляться с различными типами деградации, не получая на входе информации о том, какой именно тип искажения присутствует. Это принципиально отличается от традиционного подхода, где для каждой задачи (например,

шумоподавление и борьба с реверберацией) обучалась отдельная модель, и требовалось предварительно определить тип искажения для выбора соответствующего алгоритма.

Одним из наиболее ценных результатов соревнований URGENT Challenge стал глубокий анализ проблем, связанных с качеством обучающих данных. Эти проблемы долгое время оставались «скрытыми» в тени архитектурных достижений, но именно они часто являются первопричиной низкой точности работы моделей. Исследование выявило несколько системных проблем, первая из которых — рассогласование полосы пропускания. Во многих «высококачественных» речевых корпусах эффективная полоса пропускания не соответствует заявленной: файл, маркированный как 48 кГц, может иметь реальную верхнюю граничную частоту значительно ниже 24 кГц из-за особенностей оборудования записи или последующей обработки. Это приводит к неявному смещению моделей, т. е. они обучаются ожидать определенное частотное распределение, которое отсутствует в тестовых данных. Вторая проблема — шум разметки. Даже в эталонных корпусах, таких как DNS Challenge, присутствуют артефакты и несоответствия, негативно влияющие на обучение: «чистые» референсные сигналы могут содержать фоновые шумы, искажения микрофона или ошибки синхронизации. Модель, обучаемая на таких данных, учится не только подавлять шум, но и воспроизводить артефакты. Третья проблема — неэффективность фильтрации: простые методы фильтрации на основе популярной неинтрузивной метрики качества — Deep Noise Suppression Mean Opinion Score (DNSMOS) не удаляют эффективно шумные семплы, поскольку DNSMOS оценивает относительное качество, а не бинарную «чистоту» сигнала. Семпл с низкой громкостью или незначительными фоновыми шумами может получить низкую оценку DNSMOS, но быть вполне пригодным для обучения, тогда как семпл с артефактами кодирования имеет среднюю оценку, но вносит систематическое смещение.

Эти наблюдения привели к методологическому сдвигу в сообществе SE. Вместо стремления к «идеально чистым» данным (которые в реальности недостижимы), исследователи обратились к разработке методов, устойчивых к шуму в обучающих данных. В рамках URGENT Challenge 2025 и 2026 участникам было предложено намеренно использовать потенциально зашумленные данные и разрабатывать стратегии их эффективного использования: от методов фильтрации до методов обучения, устойчивых к шуму в метках (label-noise robust learning), и техник обучения.

Обсуждение и перспективные направления

Выполненный систематический анализ современных нейросетевых методов SE позволяет не только констатировать впечатляющий прогресс, достигнутый за последнее десятилетие, но и выявить ряд фундаментальных проблем, которые до сих пор остаются нерешенными.

Несмотря на то, что современные SE-системы демонстрируют впечатляющие результаты на стандартных бенчмарках, их практическое применение по-прежнему ограничено рядом факторов. Первая и, возможно, наиболее значимая проблема — недостаточная обобщающая способность моделей на реальные акустические условия. Модели, обученные на синтетических смесях речи и шума из ограниченного набора источников, часто показывают сильное падение качества при работе с записями, сделанными в неконтролируемых условиях. Соревнования по типу URGENT Challenge убедительно продемонстрировали, что универсальность остается недостижимой целью для большинства современных архитектур, и даже лучшие модели демонстрируют значительный разрыв в производительности между знакомыми и незнакомыми условиями.

Вторая проблема связана с перекрытием речи — ситуацией, когда одновременно говорят несколько человек. Большинство существующих SE-систем, особенно предназначенных для монокулярной обработки, практически не способны разделять перекрывающиеся голоса, что является критическим ограничением для реальных сценариев, таких как встречи, переговоры или общественные места. Даже современные архитектуры на основе трансформеров и диффузионных моделей, показывающие отличные результаты при подавлении аддитивного шума, в условиях коктейль-вечеринки часто сохраняют лишь один из голосов или смешивают их в неразборчивый шум. Эта проблема требует принципиально иных подходов, возможно, с привлечением визуальной информации (например, видео с движением губ) или многоканальной обработки.

Третья проблема — вычислительная сложность современных высококачественных моделей. Диффузионные модели, обеспечивающие наилучшую естественность звучания, требуют десятков и сотен итераций для восстановления одного сигнала, что делает их непригодными для использования в режиме реального времени. Трансформеры страдают от квадратичной сложности. Даже относительно легковесные модели могут быть слишком тяжелыми для мобильных устройств с жесткими ограничениями по энергопотреблению и задержке. Хотя архитектуры на основе Mamba и xLSTM предлагают линейную сложность, они пока не достигли качества диффузионных моделей, а их эффективность для широкого спектра искажений еще требует подтверждения.

Концепция универсального SE, заложенная URGENT Challenge, требует создания моделей, способных одновременно бороться с множеством типов искажений: шумом, реверберацией, клиппированием, ограничением полосы, артефактами кодирования, потерями пакетов и ветровым шумом. Однако текущее состояние дел таково, что большинство моделей все еще проектируются и оптимизируются под один-два типа искажений. Расширение спектра задач часто приводит к ухудшению производительности на каждом отдельном типе.

Перспективным направлением является разработка адаптивных моделей, которые могут динамически изменять свое поведение в зависимости от обнаруженного типа искажения. Также наиболее перспективными представляются три взаимосвязанных направления:

разработка легковесных архитектур (Mamba, xLSTM, дистиллированные модели) для работы на периферийных устройствах; создание универсальных моделей, способных одновременно бороться с широким спектром искажений, возможно, на основе больших предобученных аудиомоделей; интеграция SE с генеративными моделями и большими языковыми моделями для семантически-управляемого восстановления в экстремальных условиях. Параллельно должны развиваться метрики, лучше коррелирующие с человеческим восприятием, и стандартизированные протоколы сбора данных, учитывающие многообразие языков, акцентов и акустических сценариев.

Заключение

Представлен систематический обзор современных нейросетевых методов улучшения качества речевых сигналов Speech Enhancement (SE), охватывающий эволюцию архитектурных подходов. За последнее десятилетие область SE пережила трансформацию от простых полносвязных и рекуррентных сетей до сложных гибридных архитектур, сочетающих сверточные блоки, механизмы внимания, модели состояний и генеративные компоненты. Анализ эволюции позволяет выделить две ключевые тенденции: движение от монолитных архитектур к гибридным (CNN, LSTM, xLSTM, трансформеры, Mamba, Conformer, DPRNN) и переход от чисто дискриминативных к гибридным дискриминативно-генеративным системам, где дискриминативные модели обеспечивают высокое отношение сигнал/шум, но страдают от сглаживания, а генеративные диффузионные модели дают наивысшую естественность, но медленны и подвержены артефактам.

Перспективные направления дальнейших исследований включают создание легковесных моделей для мобильных устройств и слуховых аппаратов, разработку универсальных SE-систем для множества типов искажений, интеграцию SE с генеративными моделями и большими языковыми моделями для семантически-управляемого восстановления. Область SE продолжает активно развиваться, и в обозримом будущем можно ожидать создания систем, способных эффективно работать в любых акустических условиях.

Литература

1. Wang H., Wang D.L. Cross-domain diffusion based speech enhancement for very noisy speech // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. P. 1–5. doi: 10.1109/ICASSP49357.2023.10096985
2. Richter J., Welker S., Lemercier J.M., Lay B., Gerkmann T. Speech enhancement and dereverberation with diffusion-based generative models // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023. V. 31. P. 2351–2364. doi: 10.1109/TASLP.2023.3285241
3. Zaburdaev A., Ivanko D., Ryumin D. CrossMP-SENet: transformer-based cross-attention for joint magnitude-phase speech enhancement // Lecture Notes in Computer Science. 2026. V. 16188. P. 174–188. doi: 10.1007/978-3-032-07959-6_13
4. Upadhyay N., Karmakar A. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study // Procedia Computer Science. 2015. V. 54. P. 574–584. doi: 10.1016/j.procs.2015.06.066
5. Abd El-Fattah M.A., Dessouky M.I., Abbas A.M., Diab S.M., El-Rabaie E.M., Al-Nuaimy W., et al. Speech enhancement with an adaptive Wiener filter // International Journal of Speech Technology. 2014. V. 17. N 1. P. 53–64. doi: 10.1007/s10772-013-9205-5
6. Ephraim Y., van Trees H.L. A signal subspace approach for speech enhancement // IEEE Transactions on Speech and Audio Processing. 1995. V. 3. N 4. P. 251–266. doi: 10.1109/89.397090
7. Хорев А.А., Дворянкин С.В., Козлачков С.Б., Василевская Н.В. Анализ предельных возможностей методов шумопонижения и реконструкции речевых сигналов, маскируемых различными типами помех // Вопросы кибербезопасности. 2024. № 1 (59). С. 89–100. doi: 10.21681/2311-3456-2024-1-89-100
8. Pascual S., Bonafonte A., Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network // Interspeech. 2017. P. 3642–3646. doi: 10.21437/Interspeech.2017-1428
9. Fu S.-W., Tsao Y., Lu X., Kawai H. Raw waveform-based speech enhancement by fully convolutional networks // Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2017. P. 6–12. doi: 10.1109/APSIPA.2017.8281993
10. Axyonov A., Ryumin D., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition in-the-wild: Multi-angle vehicle cabin corpus and attention-based method // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. P. 8195–8199. doi: 10.1109/ICASSP48485.2024.10448048
11. Ivanko D., Karpov A., Ryumin D., Kipyatkova I., Saveliev A., Budkov V., et al. Using a high-speed video camera for robust audio-

References

1. Wang H., Wang D.L. Cross-domain diffusion based speech enhancement for very noisy speechю *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096985
2. Richter J., Welker S., Lemercier J.M., Lay B., Gerkmann T. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, vol. 31, pp. 2351–2364. doi: 10.1109/TASLP.2023.3285241
3. Zaburdaev A., Ivanko D., Ryumin D. CrossMP-SENet: transformer-based cross-attention for joint magnitude-phase speech enhancement. *Lecture Notes in Computer Science*, 2026, vol. 16188, pp. 174–188. doi: 10.1007/978-3-032-07959-6_13
4. Upadhyay N., Karmakar A. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. *Procedia Computer Science*, 2015, vol. 54, pp. 574–584. doi: 10.1016/j.procs.2015.06.066
5. Abd El-Fattah M.A., Dessouky M.I., Abbas A.M., Diab S.M., El-Rabaie E.M., Al-Nuaimy W., et al. Speech enhancement with an adaptive Wiener filter. *International Journal of Speech Technology*, 2014, vol. 17, no, 1, pp. 53–64. doi: 10.1007/s10772-013-9205-5
6. Ephraim Y., van Trees H.L. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 1995, vol. 3, no. 4, pp. 251–266. doi: 10.1109/89.397090
7. Horev A.A., Dvoryankin S.V., Kozlachkov S.B., Vasilevskaya N.V. The analysis of the potential capabilities of methods of noise reduction and reconstruction of acoustic speech signals masked by various types of noise. *Voprosy Kiberbezopasnosti*, 2024, no. 1 (59), pp. 89–100. (in Russian). doi: 10.21681/2311-3456-2024-1-89-100
8. Pascual S., Bonafonte A., Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network. *Interspeech*, 2017, pp. 3642–3646. doi: 10.21437/Interspeech.2017-1428
9. Fu S.-W., Tsao Y., Lu X., Kawai H. Raw waveform-based speech enhancement by fully convolutional networks. *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 6–12. doi: 10.1109/APSIPA.2017.8281993
10. Axyonov A., Ryumin D., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition in-the-wild: Multi-angle vehicle cabin corpus and attention-based method. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 8195–8199. doi: 10.1109/ICASSP48485.2024.10448048
11. Ivanko D., Karpov A., Ryumin D., Kipyatkova I., Saveliev A., Budkov V., et al. Using a high-speed video camera for robust audio-

visual speech recognition in acoustically noisy conditions // Lecture Notes in Computer Science. 2017. V. 10458. P. 757–766. doi: 10.1007/978-3-319-66429-3_76
12. Pandey A., Wang D.L. A new framework for CNN-based speech enhancement in the time domain // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019. V. 27. N 7. P. 1179–1188. doi: 10.1109/TASLP.2019.2913512
13. Ivanko D., Ryumin D., Axyonov A., Kashevnik A. Speaker-dependent visual command recognition in vehicle cabin: methodology and evaluation // Lecture Notes in Computer Science. 2021. P. 291–302. doi: 10.1007/978-3-030-87802-3_27
14. Pandey A., Wang D.L. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 6629–6633. doi: 10.1109/ICASSP40776.2020.9054536
15. Stoller D., Ewert S., Dixon S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation // arXiv. 2018. arXiv:1806.03185. doi: 10.48550/arXiv.1806.03185
16. Lan C., Jiang J., Zhang L., Zeng Z. Blind source separation based on improved Wave-U-Net network // IEEE Access. 2023. V. 11. P. 125951–125958. doi: 10.1109/ACCESS.2023.3330160
17. Luo Y., Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019. V. 27. N 8. P. 1256–1266. doi: 10.1109/TASLP.2019.2915167
18. Lee D., Kim S., Choi J.-W. Inter-channel Conv-TasNet for multichannel speech enhancement // arXiv. 2021. arXiv:2111.04312. doi: 10.48550/arXiv.2111.04312
19. Zheng C., Zhang H., Liu W., Luo X., Li A., Li X., et al. Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods // Trends in Hearing. 2023. V. 27. P. 1–52. doi: 10.1177/23312165231209913
20. Wang J., Saleem N., Gunawan T.S. Towards efficient recurrent architectures: A deep LSTM neural network applied to speech enhancement and recognition // Cognitive Computation. 2024. V. 16. N 3. P. 1221–1236. doi: 10.1007/s12559-024-10288-y
21. Pandey A., Wang D.L. Self-attending RNN for speech enhancement to improve cross-corpus generalization // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022. V. 30. P. 1374–1385. doi: 10.1109/taslp.2022.3161143
22. Ivanko D., Ryumin D., Kashevnik A., et al. DAVIS: Driver's Audio-Visual Speech recognition // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2022. P. 1141–1142.
23. Saleem N., Gao J., Khattak M.I., Rauf H.T., Kadry S., Shafi M. DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition // Knowledge-Based Systems. 2022. V. 238. P. 107914. doi: 10.1016/j.knosys.2021.107914
24. Аксёнов А.А., Рюмина Е.В., Рюмин Д.А., Иванько Д.В., Карпов А.А. Нейросетевой метод визуального распознавания голосовых команд водителя с использованием механизма внимания // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 4. С. 767–775. doi: 10.17586/2226-1494-2023-23-4-767-775
25. Beck M., Pöppel K., Spanring M., Auer A., Prudnikova O., Kopp M., et al. xLSTM: extended long short-term memory // Proc. of the 38th International Conference on Neural Information Processing Systems. 2024. P. 107547–107603.
26. Zhang Q., Chen M., Song Z., Liu H., Zhang X., et al. Long-context modeling networks for monaural speech enhancement: a comparative study // Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 2025. P. 1–5. doi: 10.1109/WASPAA66052.2025.11230983
27. Kühne N.L., Østergaard J., Jensen J., Tan Z.-H. xLSTM-SENet: xLSTM for single-channel speech enhancement // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2025. doi: 10.21437/interspeech.2025-108
28. Saleem N., Gunawan T.S., Kartiwi M., Nugroho B.S., Wijayanto I. NSE-CATNet: Deep neural speech enhancement using convolutional attention transformer network // IEEE Access. 2023. V. 11. P. 66979–66994. doi: 10.1109/ACCESS.2023.3290908
29. Yu W., Zhou J., Wang H.B., Tao L. SETransformer: Speech enhancement transformer // Cognitive Computation. 2022. V. 14. N 3. P. 1152–1158. doi: 10.1007/s12559-020-09817-2

visual speech recognition in acoustically noisy conditions. *Lecture Notes in Computer Science*, 2017, vol. 10458, pp. 757–766. doi: 10.1007/978-3-319-66429-3_76
12. Pandey A., Wang D.L. A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, vol. 27, no. 7, pp. 1179–1188. doi: 10.1109/TASLP.2019.2913512
13. Ivanko D., Ryumin D., Axyonov A., Kashevnik A. Speaker-dependent visual command recognition in vehicle cabin: methodology and evaluation. *Lecture Notes in Computer Science*, 2021, pp. 291–302. doi: 10.1007/978-3-030-87802-3_27
14. Pandey A., Wang D.L. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6629–6633. doi: 10.1109/ICASSP40776.2020.9054536
15. Stoller D., Ewert S., Dixon S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv*, 2018. arXiv:1806.03185. doi: 10.48550/arXiv.1806.03185
16. Lan C., Jiang J., Zhang L., Zeng Z. Blind source separation based on improved Wave-U-Net network. *IEEE Access*, 2023, vol. 11, pp. 125951–125958. doi: 10.1109/ACCESS.2023.3330160
17. Luo Y., Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, vol. 27, no. 8, pp. 1256–1266. doi: 10.1109/TASLP.2019.2915167
18. Lee D., Kim S., Choi J.-W. Inter-channel Conv-TasNet for multichannel speech enhancement. *arXiv*, 2021. arXiv:2111.04312. doi: 10.48550/arXiv.2111.04312
19. Zheng C., Zhang H., Liu W., Luo X., Li A., Li X., et al. Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods. *Trends in Hearing*, 2023, vol. 27, pp. 1–52. doi: 10.1177/23312165231209913
20. Wang J., Saleem N., Gunawan T.S. Towards efficient recurrent architectures: A deep LSTM neural network applied to speech enhancement and recognition. *Cognitive Computation*, 2024, vol. 16, no. 3, pp. 1221–1236. doi: 10.1007/s12559-024-10288-y
21. Pandey A., Wang D.L. Self-attending RNN for speech enhancement to improve cross-corpus generalization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022, vol. 30, pp. 1374–1385. doi: 10.1109/taslp.2022.3161143
22. Ivanko D., Ryumin D., Kashevnik A., et al. DAVIS: Driver's Audio-Visual Speech recognition. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2022, pp. 1141–1142.
23. Saleem N., Gao J., Khattak M.I., Rauf H.T., Kadry S., Shafi M. DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition. *Knowledge-Based Systems*, 2022, vol. 238, pp. 107914. doi: 10.1016/j.knosys.2021.107914
24. Axyonov A.A., Ryumina E.V., Ryumin D.A., Ivanko D.V., Karpov A.A. Neural network-based method for visual recognition of driver's voice commands using attention mechanism. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 767–775. (in Russian). doi: 10.17586/2226-1494-2023-23-4-767-775
25. Beck M., Pöppel K., Spanring M., Auer A., Prudnikova O., Kopp M., et al. xLSTM: extended long short-term memory. *Proc. of the 38th International Conference on Neural Information Processing Systems*, 2024, pp. 107547–107603.
26. Zhang Q., Chen M., Song Z., Liu H., Zhang X., et al. Long-context modeling networks for monaural speech enhancement: a comparative study. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5. doi: 10.1109/WASPAA66052.2025.11230983
27. Kühne N.L., Østergaard J., Jensen J., Tan Z.-H. xLSTM-SENet: xLSTM for single-channel speech enhancemen. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2025, doi: 10.21437/interspeech.2025-108
28. Saleem N., Gunawan T.S., Kartiwi M., Nugroho B.S., Wijayanto I. NSE-CATNet: Deep neural speech enhancement using convolutional attention transformer network. *IEEE Access*, 2023, vol. 11, pp. 66979–66994. doi: 10.1109/ACCESS.2023.3290908
29. Yu W., Zhou J., Wang H.B., Tao L. SETransformer: Speech enhancement transformer. *Cognitive Computation*, 2022, vol. 14, no. 3, pp. 1152–1158. doi: 10.1007/s12559-020-09817-2

30. Jannu C., Vanambathina S.D. Convolutional transformer based local and global feature learning for speech enhancement // International Journal of Advanced Computer Science and Applications. 2023. V. 14. N 1. P. 81. doi: 10.14569/IJACSA.2023.0140181
31. Лепендин А.А., Насретдинов Р.С., Ильяшенко И.Д. Метод улучшения качества речи с использованием модифицированного кодирующего-декодирующего пирамидального трансформера // Труды Института системного программирования РАН. 2022. Т. 34. № 4. С. 135–152. doi: 10.15514/ISPRAS-2022-34(4)-10
32. Li M., Liu Y., Zhou L. DeConformer-SENet: An efficient deformable conformer speech enhancement network // Digital Signal Processing. 2025. V. 156. Part A. P. 104787. doi: 10.1016/j.dsp.2024.104787
33. Xu X., Tu W., Yang Y. Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2023. doi: 10.21437/interspeech.2023-1376
34. Rekesh D., Koluguri N.R., Kriman S., Majumdar S., Noroozi V., Huang H., et al. Fast conformer with linearly scalable attention for efficient speech recognition // Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2023. P. 1–8. doi: 10.1109/asru57964.2023.10389701
35. Gu A., Dao T. Mamba: Linear-time sequence modeling with selective state spaces // arXiv. 2023. arXiv:2312.00752. doi: 10.48550/arXiv.2312.00752
36. Chao R., Cheng W.-H., La Quatra M., Siniscalchi S.M., Huck Yang C.-H., Fu S.-W., et al. An investigation of incorporating mamba for speech enhancement // Proc. of the IEEE Spoken Language Technology Workshop (SLT). 2024. P. 302–308. doi: 10.1109/slt61566.2024.10832332
37. Wang J., Lin Z., Wang T., Ge M., Wang L., Dang J. Mamba-SEUNet: Mamba UNet for monaural speech enhancement // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025. P. 1–5. doi: 10.1109/ICASSP49660.2025.10889525
38. Chen M., Zhang Q., Wang M., Zhang X., Liu H., Ambikairaiah E., et al. Selective state space model for monaural speech enhancement // IEEE Transactions on Consumer Electronics. 2025. V. 71. N 2. P. 5414–5424. doi: 10.1109/TCE.2024.3523297
39. Zhang X., Zhang Q., Liu H., Xiao T., Qian X., Ahmed B., et al. Mamba in speech: Towards an alternative to self-attention // IEEE Transactions on Audio, Speech and Language Processing. 2025. V. 33. P. 1933–1948. doi: 10.1109/TASLPRO.2025.3566210
40. Chung H., Plourde E., Champagne B. Discriminative training of NMF model based on class probabilities for speech enhancement // IEEE Signal Processing Letters. 2016. V. 23. N 4. P. 502–506. doi: 10.1109/LSP.2016.2532903
41. Li C., Cornell S., Watanabe S., Qian Y. Diffusion-based generative modeling with discriminative guidance for streamable speech enhancement // Proc. of the IEEE Spoken Language Technology Workshop (SLT). 2024. P. 333–340. doi: 10.1109/SLT61566.2024.10832147
42. Donahue C., Li B., Prabhavalkar R. Exploring speech enhancement with generative adversarial networks for robust speech recognition // Proc. of the IEEE international conference on acoustics, speech and signal processing (ICASSP). 2018. P. 5024–5028. doi: 10.1109/ICASSP.2018.8462581
43. Lu Y.J., Wang Z.-Q., Watanabe S., Richard A., Yu C., Tsao Y. Conditional diffusion probabilistic model for speech enhancement // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. P. 7402–7406. doi: 10.1109/ICASSP43922.2022.9746901
44. Yen H., Germain F.G., Wichern G., Le Roux J. Cold diffusion for speech enhancement // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. P. 1–5. doi: 10.1109/ICASSP49357.2023.10096064
45. Serrà J., Pascual S., Pons J., Oguz Araz R., Scaini D. Universal speech enhancement with score-based diffusion // arXiv. 2022. arXiv:2206.03065. doi: 10.48550/arXiv.2206.03065
46. Lu Y.-J., Tsao Y., Watanabe S. A study on speech enhancement based on diffusion probabilistic model // Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2021. P. 659–666.
47. Saijo K., Zhang W., Cornell S., Scheibler R., Li C., Ni Z., et al. Interspeech 2025 URGENT speech enhancement challenge // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2025. doi: 10.21437/interspeech.2025-1363

30. Jannu C., Vanambathina S.D. Convolutional transformer based local and global feature learning for speech enhancement. International *Journal of Advanced Computer Science and Applications*, 2023, vol. 14, no. 1, pp. 81. doi: 10.14569/IJACSA.2023.0140181
31. Lependin A.A., Nasretdinov R.S., Ilyashenko I.D. Speech enhancement method based on modified encoder-decoder pyramid transformer. *Proceedings of the Institute for System Programming of the RAS*, 2022, vol. 34, no. 4, pp. 135–152. (in Russian). doi: 10.15514/ISPRAS-2022-34(4)-10
32. Li M., Liu Y., Zhou L. DeConformer-SENet: An efficient deformable conformer speech enhancement network. *Digital Signal Processing*, 2025, vol. 156, part A, pp. 104787. doi: 10.1016/j.dsp.2024.104787
33. Xu X., Tu W., Yang Y. Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2023, doi: 10.21437/interspeech.2023-1376
34. Rekesh D., Koluguri N.R., Kriman S., Majumdar S., Noroozi V., Huang H., et al. Fast conformer with linearly scalable attention for efficient speech recognition. *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8. doi: 10.1109/asru57964.2023.10389701
35. Gu A., Dao T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv*, 2023. arXiv:2312.00752. doi: 10.48550/arXiv.2312.00752
36. Chao R., Cheng W.-H., La Quatra M., Siniscalchi S.M., Huck Yang C.-H., Fu S.-W., et al. An investigation of incorporating mamba for speech enhancement. *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 302–308. doi: 10.1109/slt61566.2024.10832332
37. Wang J., Lin Z., Wang T., Ge M., Wang L., Dang J. Mamba-SEUNet: Mamba UNet for monaural speech enhancement. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10889525
38. Chen M., Zhang Q., Wang M., Zhang X., Liu H., Ambikairaiah E., et al. Selective state space model for monaural speech enhancement. *IEEE Transactions on Consumer Electronics*, 2025, vol. 71, no. 2, pp. 5414–5424. doi: 10.1109/TCE.2024.3523297
39. Zhang X., Zhang Q., Liu H., Xiao T., Qian X., Ahmed B., et al. Mamba in speech: Towards an alternative to self-attention. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, vol. 33, pp. 1933–1948. doi: 10.1109/TASLPRO.2025.3566210
40. Chung H., Plourde E., Champagne B. Discriminative training of NMF model based on class probabilities for speech enhancement. *IEEE Signal Processing Letters*, 2016, vol. 23, no. 4, pp. 502–506. doi: 10.1109/LSP.2016.2532903
41. Li C., Cornell S., Watanabe S., Qian Y. Diffusion-based generative modeling with discriminative guidance for streamable speech enhancement. *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 333–340. doi: 10.1109/SLT61566.2024.10832147
42. Donahue C., Li B., Prabhavalkar R. Exploring speech enhancement with generative adversarial networks for robust speech recognition. *Proc. of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018, pp. 5024–5028. doi: 10.1109/ICASSP.2018.8462581
43. Lu Y.J., Wang Z.-Q., Watanabe S., Richard A., Yu C., Tsao Y. Conditional diffusion probabilistic model for speech enhancement. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7402–7406. doi: 10.1109/ICASSP43922.2022.9746901
44. Yen H., Germain F.G., Wichern G., Le Roux J. Cold diffusion for speech enhancement. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096064
45. Serrà J., Pascual S., Pons J., Oguz Araz R., Scaini D. Universal speech enhancement with score-based diffusion. *arXiv*, 2022. arXiv:2206.03065. doi: 10.48550/arXiv.2206.03065
46. Lu Y.-J., Tsao Y., Watanabe S. A study on speech enhancement based on diffusion probabilistic model. Proc. of the Asia-Pacific Signal and *Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 659–666.
47. Saijo K., Zhang W., Cornell S., Scheibler R., Li C., Ni Z., et al. Interspeech 2025 URGENT speech enhancement challenge. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2025, doi: 10.21437/interspeech.2025-1363

48. Dubey H., Aazami A., Gopal V., Naderi B., Braun S., Cutler R., et al. ICASSP 2023 deep noise suppression challenge // IEEE Open Journal of Signal Processing. 2024. V. 5. P. 725–737. doi: 10.1109/OJSP.2024.3378602
49. Watanabe S., Mandel M., Barker J., Vincent E., Arora A., Chang X., et al. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings // arXiv. 2020. arXiv:2004.09249. doi: 10.48550/arXiv.2004.09249

48. Dubey H., Aazami A., Gopal V., Naderi B., Braun S., Cutler R., et al. ICASSP 2023 deep noise suppression challenge. *IEEE Open Journal of Signal Processing*, 2024, vol. 5, pp. 725–737. doi: 10.1109/OJSP.2024.3378602
49. Watanabe S., Mandel M., Barker J., Vincent E., Arora A., Chang X., et al. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv*, 2020. arXiv:2004.09249. doi: 10.48550/arXiv.2004.09249

Автор

Иванько Денис Викторович — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57190967993, https://orcid.org/0000-0003-0412-7765, denis.ivanko11@gmail.com

Author

Denis V. Ivanko — PhD, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57190967993, https://orcid.org/0000-0003-0412-7765, denis.ivanko11@gmail.com

Статья поступила в редакцию 20.04.2026
Одобрена после рецензирования 29.04.2026
Принята к печати 25.07.2026

Received 20.04.2026
Approved after reviewing 29.04.2026
Accepted 25.07.2026