Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

на правах рукописи

Андреев Павел Константинович

ГЕНЕРАТИВНЫЕ МОДЕЛИ ДЛЯ УЛУЧШЕНИЯ РЕЧИ

РЕЗЮМЕ

диссертации на соискание ученой степени

кандидата компьютерных наук

Москва — 2024

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель: Конушин Антон Сергеевич, к.ф.-м.н., «Национальный исследовательский университет «Высшая школа экономики».

1 Введение

1.1 Тема работы

Записи речи часто содержат фоновый шум, реверберацию, снижение частотной полосы и другие искажения, что снижает как их разборчивость, так и эстетическое удовлетворение слушателя. Технологии улучшения речи предназначены для восстановления чистой речи из таких искаженных сигналов. Технологии улучшения речи снижают технические требования к оборудованию для записи, позволяя профессионалам создавать записи студийного качества без сложного студийного оборудования. Модели улучшения речи могут использоваться для обеспечения коммуникации в акустически загрязненных средах и особенно важны для технологий помощи лицам с нарушениями слуха. Кроме того, модели улучшения речи играют критическую роль в создании высококачественных наборов данных для систем глубокого обучения, которые требуют обширных наборов данных чистой речи для эффективного обучения. Таким образом, общедоступные данные, независимо от исходных условий записи, могут быть улучшены до качества, необходимого для передовых моделей синтеза речи. Следовательно, технологии улучшения речи важны для широкого спектра приложений.

Более формальное определение задачи улучшения речи может быть выражено в вероятностной формулировке. Пусть $p(y)$ — распределение чистой речи, а $p(x|y)$ — модель деградации. Задача улучшения речи заключается в восстановлении речи из условного распределения $p(y|x)$, где $x \sim p(x|y)$ является сигналом речи, искаженным моделью деградации. Распределение модели деградации $p(x|y)$ может включать различные формы преобразований сигнала, включая добавление фонового шума, снижение частотной полосы, артефакты кодека, реверберацию и так далее. В зависимости от доступных ресурсов и конкретного применения можно выделить несколько формулировок задачи улучшения речи:

- **Базовое улучшение речи** [48, 32]: Этот сценарий предполагает улучшение речи без ограничений по задержке. В данной формулировке распределение деградаций известно заранее, и модель не ограничена требованием потоковой обработки.

- **Потоковое улучшение речи** [19, 10]: В этом сценарии требуется создание моделей улучшения речи с ограниченной алгоритмической задержкой, т.е. каузальных моделей. Потоковое улучшение речи накладывает ограничения на использование модели только ограниченного окна будущей информации, обычно от 3-8 мс (низкая задержка) до 60 мс (высокая задержка).

- **Улучшение речи без учителя** [30, 35]: Эта формулировка не предполагает, что распределение модели деградации $p(x|y)$ известно заранее. Модель адаптируется к конкретной модели деградации только на этапе предсказаний.

Разработка эффективных генеративных моделей для улучшения речи в контексте всех этих формулировок является **темой данной работы**.

1.2 Актуальность

Исторически, в области обработки аудио превалировали методы, основанные на вручную разработанных эвристиках и статистических моделях, часто использующих нереалистичные предположения о структуре речи и помехах [14, 13]. Однако с развитием машинного обучения и, в частности, глубокого обучения произошел сдвиг парадигмы в сторону использования методов, основанных на данных. В отличие от традиционных подходов, парадигма, основанная на данных, изучает характеристики сигналов непосредственно из данных. Этот подход оказался полезным во многих областях, включая обработку речи.

Первые попытки применения методов глубокого обучения к задаче улучшения речи основывались на трактовке этой задачи как предсказательной [10, 16, 5, 19]. Следуя принципу минимизации эмпирического риска, целью предсказательного моделирования является нахождение модели с минимальной средней ошибкой на обучающих данных. Получив шумный звуковой сигнал или спектрограмму, эти подходы пытаются предсказать чистый сигнал, минимизируя точечное расстояние в доменах времени, спектра, или совместно в обоих доменах, таким образом трактуя эту задачу как предсказательную.

Однако при наличии значительных искажений сигнала существует неопределенность в восстановлении речевого сигнала (т.е. при данном искаженном сигнале чистый сигнал не восстанавливается однозначно), что часто приводит к чрезмерному сглаживанию предсказанной речи. С вероятностной точки зрения минимизация точечного расстояния приводит к эффекту усреднения. Например, оптимизация среднеквадратичной ошибки между звуковыми сигналами дает математическое ожидание сигнала по условному распределению чистой речи, обусловленному на ее искаженный вариант. Ключевая проблема заключается в том, что математическое ожидание по этому распределению может не лежать в пределах этого распределения.

На Рис. 1 показан наглядный пример этого явления и его влияния на улучшение речи. Модель обучалась расширять частотную полосу речевого сигнала, обуславливаясь на сигнал с уменьшенной полосой частот, минимизируя среднеквадратичную ошибку между чистыми и генериру-

мыми звуковыми сигналами [3]. Примечательно, что модель не способна восстановить высокие частоты чистого сигнала при минимизации среднеквадратичной ошибки. Из-за высокой неопределенности модель чрезмерно сглаживает высокие частоты, не в состоянии восстановить речевой сигнал выше 5 кГц. Схожие эффекты наблюдаются при использовании других точечных функций потерь, включая спектральные потери.

В отличие от предсказательных моделей, генеративные модели стремятся к сэмплированию из распределения чистой речи, обусловленного на искаженный сигнал, а не к минимизации точечных потерь. Преимущество такого подхода заключается в том, что модель улучшения речи вынуждена генерировать сигнал, лежащий в пределах распределения чистой речи, как показано на рис. 1. **Данная работа изучает генеративные модели в контексте улучшения речи, предлагает новые методы и повышает эффективность и качество существующих решений.**

В первой части этой работы мы сосредоточимся на разработке эффективных решений для базового улучшения речи [48, 32]. Формулировка базового улучшения речи не накладывает ограничений на доступный контекст сигнала, и модель деградации считается известной на этапе обучения. Мы, во-первых, утверждаем, что задача улучшения речи не требует от модели выучивания полного условного распределения $p(y|x)$, вместо этого мы утверждаем необходимость предсказания главной моды этого распределения.

Мы показываем, что структура обучения генеративно-состязательных сетей (GAN) естественно подходит для предложенной формулировки задачи улучшения речи, поскольку данная структура позволяет восстановить основную моду условного распределения — именно то, что обычно должны делать модели улучшения речи. Поэтому в этой работе мы используем структуру GAN для решения задачи базового улучшения речи и разрабатываем эффективные архитектуры генератора и дискриминатора нейронных сетей.

Во второй части работы мы сосредоточимся на потоковом улучшении речи [19, 10, 45, 51], в частности на улучшении речи с низкой задержкой [45, 51]. Потоковые модели являются важным компонентом инструментов для улучшения речи в реальном времени. Потоковой режим накладывает ограничения на использование моделями улучшения речи только небольшого контекста будущей информации. В результате, улучшение речи в режиме потоковой обработки с низкой задержкой обычно считается сложной задачей, так как данный режим оказывает значительное негативное влияние на качество модели.

Тем не менее, последовательная природа потоковой генерации предлагает естественную возможность для авторегрессии, то есть использования предыдущих предсказаний при создании текущих. Традиционным мето-

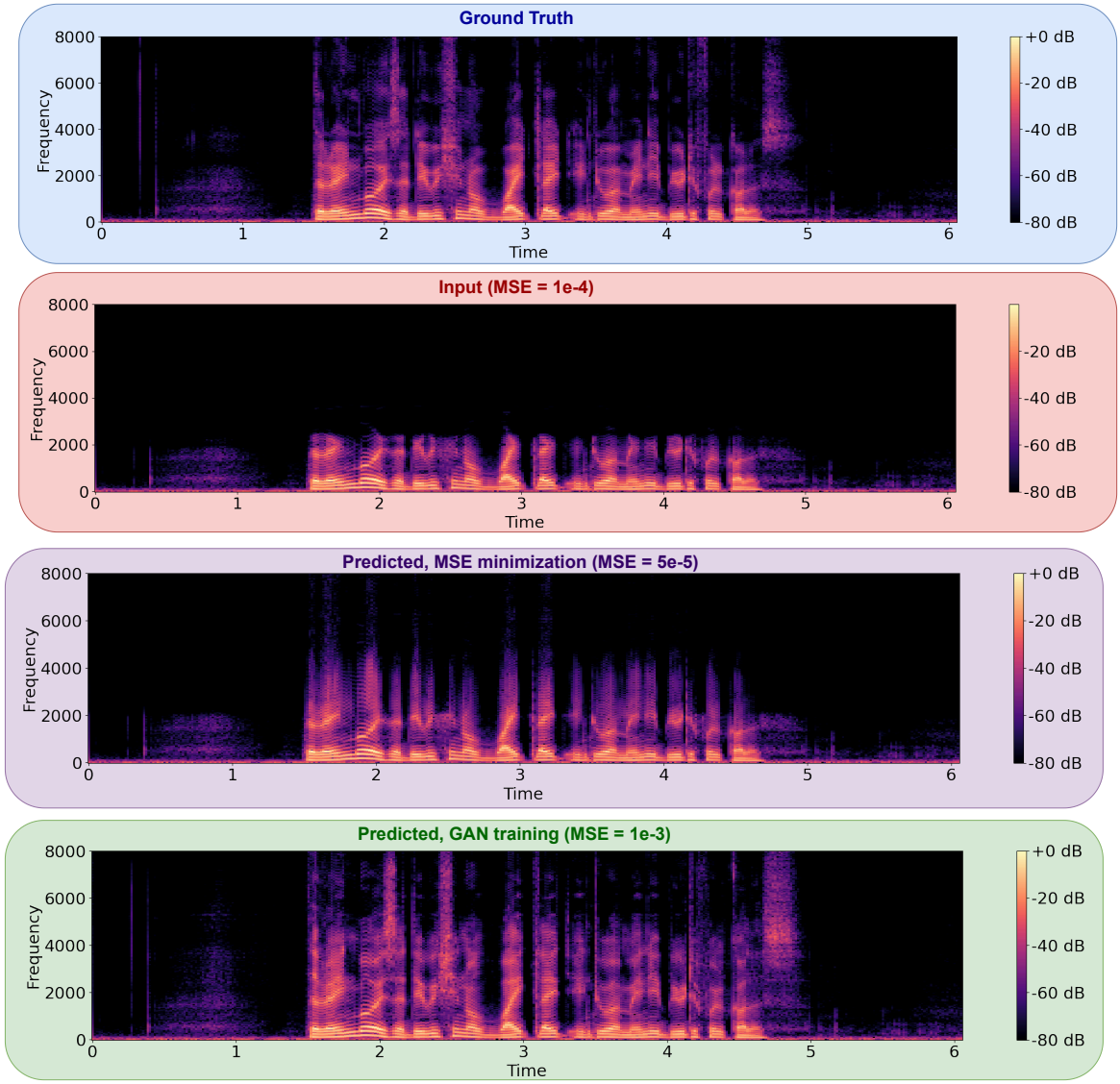

Рис. 1: Пример спектрограмм речевого сигнала. Минимизация среднеквадратичной ошибки приводит к чрезмерному сглаживанию предсказанного сигнала, что приводит к отсутствию высоких частот в предсказанном сигнале. Хотя предсказание, предоставленное генеративной моделью, приводит к среднеквадратичной ошибке, которая даже выше, чем у входного сигнала, предсказанный сигнал восстанавливает исходную речь. Аналогичный эффект наблюдается в моделях увеличивающих разрешение изображений [27].

дом обучения авторегрессионных генеративных моделей является принуждение учителя (англ. teacher forcing), но его основной недостаток заключается в несоответствии обучения и вывода, что может привести к значительному ухудшению качества. Мы предлагаем простую, но эффективную альтернативную технику обучения авторегрессионных моделей для

улучшения речи с низкой задержкой. Мы демонстрируем, что предлагаемый подход приводит к стабильному улучшению во многих архитектурах и сценариях обучения.

В конце мы сосредоточимся на проблеме улучшения речи без учителя [30, 35]. В этой части работы, мы описываем диффузионную вероятностную модель, способную решать различные задачи инверсии деградаций речевых сигналов с неизвестными моделями деградации на этапе обучения. После обучения для генерации речевого сигнала в безусловном режиме она может быть адаптирована для различных задач, включая инверсию деградаций и нейронное вокодирование.

1.3 Основные результаты и выводы

1.3.1 Вклад

Основные вклады данной работы можно резюмировать следующим образом:

1. Предложена архитектура генератора HiFi++, объединяющая генератор HiFi-GAN с тремя новыми модулями: SpectralUnet, WaveUNet и SpectralMaskNet. Эта новая архитектура генератора позволяет создать модель для расширения полосы пропускания и улучшения речи, достигая передовых результатов в этих задачах.

2. Разработана новая архитектура для непосредственной оценки комплекснозначных спектрограмм на основе оператора свертки над быстрым преобразованием Фурье. Архитектура позволяет напрямую работать с признаками кепстра и дополнительно улучшает результаты HiFi++ в задачах улучшения речи, при этом будучи более эффективной по параметрам.

3. Были исследованы различные модели создания признаков для функции потерь при генерации речи и введены критерии для выбора модели на основе структуры его пространства признаков. Эффективность этих критериев подтверждена эмпирическими результатами.

4. Основываясь на этих разработках, была создана новая универсальная модель улучшения речи FINALLY, которая достигает передовых результатов, превосходя все существующие решения, при этом являясь более вычислительно эффективной.

5. Предложен новый метод обучения авторегрессионных моделей для улучшения речи с низкой задержкой. Метод позволяет смягчить рассогласование между обучением и этапом предсказаний, возникающее

при обучении с принуждением учителя. Модель обеспечивает значительное улучшение потоковых моделей улучшения речи с авторегрессионным условием.

6. Исследована техника на основе диффузии для улучшения речи без учителя. Предложенная безусловная диффузионная модель может быть обучена для задачи безусловной генерации речи, а затем адаптирована для различных задач восстановления речи без дополнительного обучения.

1.3.2 Теоретическая и практическая значимость

В данной работе теоретически показано, что состязательное обучение может использоваться для неявной регрессии к основной моде распределения, что делает его подходящим инструментом для обучения моделей улучшения речи. Также изучены структурные свойства различных пространств моделей извлечения признаков речи и сформулирована новая функция потерь для генерации речи.

Кроме того, в работе предложены новые нейронные архитектуры, HiFi++ и FFC-SE, для глубоких генеративных моделей, улучшающих качество и вычислительную эффективность решений по улучшению речи. На основе этих разработок предложен высокоэффективный алгоритм улучшения речи FINALLY, достигающий передового качества при значительно меньших вычислительных ресурсах по сравнению с предыдущими методами.

В работе также представлен новый метод обучения авторегрессионных моделей в условиях предотвращающий рассогласование между этапами обучения и предсказаний, метод значительно превосходит традиционную технику обучения с принуждением учителя. Предлагаемый метод итеративной авторегрессии обладает значительным потенциалом для практической пользы ввиду широкого использования авторегрессионных моделей.

Также, в работе изучена проблема улучшения речи без учителя и предложена новая диффузионная генеративная модель Undiff для восстановления речи без учителя.

1.3.3 Ключевые аспекты/идеи для защиты

1. Архитектура HiFi++ для мулти-доменной обработи сигнала при улучшении речи.

2. FFC-SE для непосредственной оценки комплекснозначных спектрограмм при улучшении речи.

3. Модель FINALLY для универсального улучшения речи.

4. Итеративная авторегрессия для уменьшения эффекта рассогласования между этапами обучения и предсказаний в авторегрессионных моделях, изученная в применении к улучшению речи с низкой задержкой.

5. Диффузионная вероятностная модель Undiff для улучшения речи без учителя.

1.3.4 Личный вклад

Идея архитектур генераторов HiFi++ и FFC-SE была предложена автором этой работы. Начальная реализация архитектуры HiFi++ была выполнена автором, а Айбек Аланов и Олег Иванов помогли подготовить кодовую базу для экспериментов. Сеть FFC-SE была совместно разработана с Иваном Щекотовым. Эксперименты по проверке эффективности сетей были спроектированы автором. Реализация и написание статей выполнены совместно с Айбеком Алановым, Олегом Ивановым и Иваном Щекотовым. Научное руководство для этих работ предоставил Дмитрий Ветров.

Доказательство, что обучение LS-GAN приводит к регрессии на моду распределения и формулировка задачи улучшения речи как задачи поиска моды были разработаны автором диссертации. Модель FINALLY была разработана совместно с Кириллом Тамогашевым и Николаем Бабаевым. Автор отвечал за научное руководство, планирование экспериментов и валидацию кода.

Итеративная авторегрессия была предложена и теоретически изучена автором этой работы. Фактическая реализация и экспериментальная проверка выполнены совместно с Николаем Бабаевым. Статья была написана автором при некоторой помощи Николая Бабаева и Айбека Аланова.

Диффузионная генеративная модель Undiff была спроектирована и реализована совместно с Анастасией Ященко и Иваном Щекотовым. Научное руководство для этой работы предоставил Дмитрий Ветров.

1.4 Публикации и апробация работы

Результаты данной диссертации опубликованы в 3 публикациях повышенного уровня и 1 публикации стандартного уровня. Кандидат является основным автором всех этих статей[1].

[1] * указывает на равный вклад

1.4.1 Публикации повышенного уровня

- **Andreev, P.***, Babaev, N.*, Saginbaev, A., Shchekotov, I., Alanov, A. (2023). Iterative autoregression: a novel trick to improve your low-latency speech enhancement model. *Proc. INTERSPEECH 2023*, 2448-2452, doi: 10.21437/Interspeech.2023-365 (Core A)

- Shchekotov, I.*, **Andreev, P.***, Ivanov, O., Alanov, A., Vetrov, D. (2022). FFC-SE: Fast Fourier Convolution for Speech Enhancement. *Proc. INTERSPEECH 2022*, 1188-1192, doi: 10.21437/Interspeech.2022-603 (Core A)

- Iashchenko, A.*, **Andreev, P.***, Shchekotov, I.*, Babaev, N., Vetrov, D. (2023). UnDiff: Unsupervised Voice Restoration with Unconditional Diffusion Model. *Proc. INTERSPEECH 2023*, 4294-4298, doi: 10.21437/Interspeech.2023-367 (Core A)

1.4.2 Публикации стандартного уровня

- **P. Andreev***, A. Alanov, O. Ivanov* and D. Vetrov, "HIFI++: A Unified Framework for Bandwidth Extension and Speech Enhancement,"*ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10097255. (Core B)

1.4.3 Доклады на научных конференциях

- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, June 8, 2023. Topic: "HIFI++: A Unified Framework for Bandwidth Extension and Speech Enhancement"

- 24th INTERSPEECH Conference, Dublin, Ireland, August 22, 2023. Topic: "Iterative autoregression: a novel trick to improve your low-latency speech enhancement model"

- 24th INTERSPEECH Conference, Dublin, Ireland, August 23, 2023. Topic: "UnDiff: Unsupervised Voice Restoration with Unconditional Diffusion Model"

- 23rd INTERSPEECH Conference, Incheon, Korea, September 20, 2022. Topic: "FFC-SE: Fast Fourier Convolution for Speech Enhancement"

1.4.4 Объем и структура работы

Диссертация включает введение, в котором формулируется тема работы, обзорный раздел, представляющий необходимый контекст, три главы ос-

новного содержания, описывающие разработанные подходы для каждой из предложенных формулировок, и заключение, которое подводит итоги разработок и завершает исследование. Общий объем диссертации составляет 102 страницы.

2 Содержание работы

В главе "Введение" мы представляем основные модели деградации и метрики. Также мы делаем обзор литературы, описывающей подходы к улучшению речи на основе глубокого обучения, изложенные в предыдущих работах.

В трех последующих главах мы описываем разработанные методы для каждой из представленных формулировок задачи улучшения речи:

1. **Базовое улучшение речи** рассматривается в третьей главе, "Генеративные модели для базового улучшения речи". В этой главе сначала уточняется вероятностная формулировка задачи улучшения речи, с акцентом на ее практическую цель. Мы утверждаем, что модель улучшения речи должна восстанавливать наиболее вероятную реконструкцию чистой речи, исходя из деградированной версии. В соответствии с уточненной формулировкой, мы показываем, что обучение на основе GAN естественным образом подходит для этой цели, поощряя генератор восстанавливать моду условного распределения. Затем в главе представлены практические разработки архитектуры нейронных сетей и функций потерь для направления обучения к правильным решениям.

2. **Потоковое улучшение речи** рассматривается в четвертой главе, "Итеративная авторегрессия для потокового улучшения речи". В этой главе мы утверждаем, что потоковый режим предоставляет естественную возможность для авторегрессионного моделирования улучшения речи. Мы показываем, что традиционный алгоритм обучения с принуждением учителя приводит к значительному расхождению между обучением и выводом и представляем новый алгоритм обучения, который устраняет эту проблему.

3. **Улучшение речи без учителя** рассматривается в пятой главе, "Улучшение речи без учителя с использованием безусловной диффузионной модели". В этой главе представлена UnDiff, диффузионная вероятностная модель, способная решать различные задачи восстановления речи. После обучения для задаче безусловной генерации

речевых сигналов, она может быть адаптирована к различным задачам восстановления речи без дополнительного обучения. Мы сначала решаем задачу безусловного генерации речевых сигналов, сравнивая различные нейронные архитектуры и домены предварительной обработки. Затем мы демонстрируем, как обученная безусловная диффузионная модель может быть адаптирована к различным задачам обработки речи без дополнительного обучения.

2.1 Генеративные модели для базового улучшения речи

2.1.1 GAN для улучшения речи

Для начала конкретизируем вероятностную формулировку задачи улучшения речи, более детально рассматривая ее практическую цель. Практическая цель модели улучшения речи заключается в восстановлении чистого аудиосигнала, содержащего характеристики оригинальной записи речи, включая голос, лингвистическое содержание и просодию. Таким образом, модель улучшения речи не должна генерировать новую речь, а скорее восстановить существующую речь, как если бы она была записана в идеальных условиях (качество студийной записи). С математической точки зрения это означает, что модель улучшения речи должна восстанавливать наиболее вероятную реконструкцию чистой речи y, исходя из деградированной версии x, то есть $y = \arg\max_y p_{\text{clean}}(y|x)$.

Исходя из данной формулировки, мы утверждаем, что структура обучения генеративных состязательных сетей (GAN) естественно подходит для задачи улучшения речи. Мы показываем, что обучение GAN приводит к поиску моды распределения генератора, что соответствует введенной выше постановке.

Пусть $p_g(y|x)$ — это семейство распределений формы волны, создаваемых генератором $g_\theta(x)$. Мао и др. [33] показали, что обучение с Least Squares GAN (LS-GAN) приводит к минимизации χ^2 дивергенции Пирсона $\chi^2_{\text{Pearson}}(p_g \| (p_{\text{clean}} + p_g)/2)$. Мы показываем, что если $p_g(y|x)$ приближается к $\delta(y - g_\theta(x))$ при некоторой параметризации, то минимизация этой дивергенции приводит к $g_\theta(x) = \arg\max_y p_{\text{clean}}(y|x)$. Это означает, что если генератор детерминированно предсказывает чистую форму волны из деградированного сигнала, то потеря LS-GAN побуждает генератор предсказывать точку максимальной плотности $p_{\text{clean}}(y|x)$. Мы отмечаем, что хотя в предыдущих работах [29] было продемонстрировано свойство покрытия мод при оптимизации χ^2 дивергенции Пирсона, наш результат относится к детерминированным предсказаниям генератора, что выходит за рамки анализа, представленного Ли и др. [29].

Лемма 1. *Пусть* $p_{clean}(y|x) > 0$ *— конечная и липшицева непрерывная функция плотности с уникальным глобальным максимумом, и* $p_g^\xi(y|x) = \xi^n/2^n \cdot \mathbf{1}_{y-g_\theta(x)\in[-1/\xi,1/\xi]^n}$*, тогда*

$$\lim_{\xi\to+\infty} \arg\min_{g_\theta(x)} \chi^2_{Pearson}(p_g^\xi || (p_{clean} + p_g^\xi)/2) = \arg\max_y p_{clean}(y|x) \qquad (1)$$

Таким образом, обучение LS-GAN в идеальных условиях должно привести к решению $g_\theta(x) = \arg\max_y p_{\text{clean}}(y|x)$ для генератора. На практике, однако, успех в значительной степени зависит от технических деталей, таких как дополнительные функции потерь для стабилизации обучения и архитектуры нейронных сетей. В следующих разделах мы решаем эти проблемы, разрабатывая новые архитектуры и функции потерь для обучения.

2.1.2 HiFi++: Единая модель для расширения частотной полосы пропускания и улучшения речи

Мы предлагаем новую архитектуру HiFi++, которая адаптирует генератор HiFi [21] для задачи улучшения речи, вводя новые модули: SpectralUNet, WaveUNet и SpectralMaskNet (см. Рисунок 2). Генератор HiFi++ основан на HiFi генераторе, который принимает на вход обработанное представление мел-спектрограммы от SpectralUNet, а его выход проходит через модули постобработки: WaveUNet корректирует выходной сигнал во временной области, в то время как SpectralMaskNet очищает его в частотной области. В следующих параграфах мы подробно описываем введенные модули.

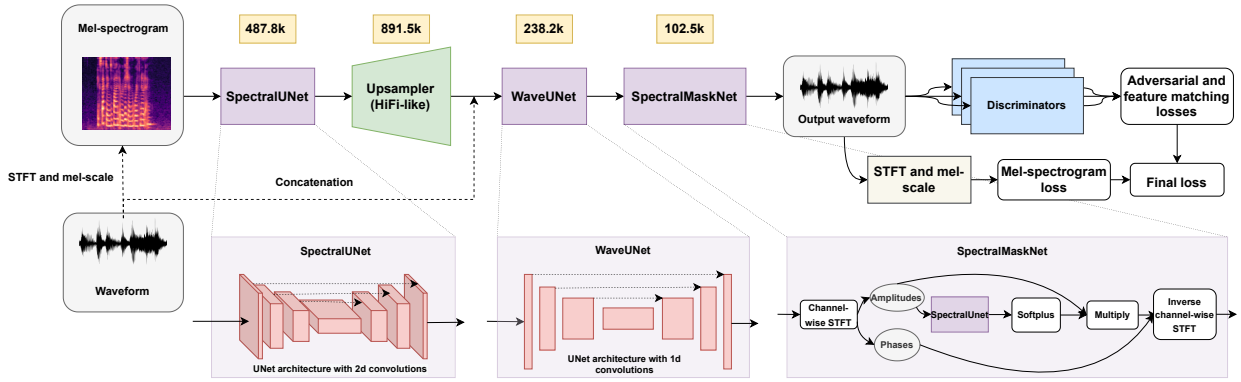

Рис. 2: Архитектура HiFi++ и процесс обучения. Генератор HiFi++ состоит из Upsampler, аналогичного HiFi генератору, и трех введенных модулей: SpectralUNet, WaveUNet и SpectralMaskNet (их размеры указаны в желтых блоках).

SpectralUNet Мы вводим модуль SpectralUNet как начальную часть генератора HiFi++, который принимает на вход мел-спектрограмму (см. Figure 2). Мел-спектрограмма имеет двумерную структуру, и двумерные сверточные блоки модели SpectralUnet разработаны для работы с этой структурой на начальном этапе преобразования мел-спектрограммы в форму волны. Идея заключается в упрощении задачи для оставшейся части генератора HiFi++, которая должна преобразовать это двумерное представление в одномерную последовательность. Модуль SpectralUNet разработан как архитектура типа UNet с двумерными свертками. Этот модуль также можно рассматривать как часть предварительной обработки, которая подготавливает входную мел-спектрограмму, исправляя и извлекая из нее важную информацию, необходимую для выполнения поставленной задачи.

WaveUNet Модуль WaveUNet размещен после части HiFi генератора (Upsampler-a) и принимает несколько одномерных последовательностей, объединенных с входной волновой формой. Этот модуль работает непосредственно во временной области и может рассматриваться как механизм постобработки во временной области, который улучшает выход Upsampler-a и объединяет предсказанную форму волны с исходной. Модуль WaveUNet сделана на основн известной архитектуры Wave-U-Net [42], которая представляет собой полностью сверточную одномерную UNet-подобную нейронную сеть. Этот модуль выводит двумерный тензор, состоящий из m одномерных последовательностей, которые будут обработаны и объединены в выходную волновую форму следующим модулем SpectralMaskNet.

SpectralMaskNet Модуль SpectralMaskNet является заключительной частью генератора, представляющую собой обучаемое спектральное маскирование. Он принимает на вход двумерный тензор из m одномерных последовательностей и применяет поканальное оконное преобразование Фурье (STFT) к этому двумерному тензору. Далее сеть, подобная SpectralUNet, принимает амплитуды выхода STFT, чтобы предсказать мультипликативные коэффициенты для этих амплитуд. Заключительная часть состоит из обратного STFT преобразования для модифицированного спектра (см. Рис. 2). Важно, что этот процесс не изменяет фазы. Цель этого модуля - выполнить постобработку сигнала в частотной области. SpectralMaskNet - это эффективный механизм для удаления артефактов и шума в частотной области из выходной волновой формы обучаемым способом.

Обучение Для обучения моделей во временной области мы используем мултидискриминаторное состязательное обучение. Оно состоит из трех

функций потерь: LS-GAN потери $\mathcal{L}_{GAN}$ [33], потери сопоставления признаков $\mathcal{L}_{FM}$ [25, 24] и потери восстановления мел-спектрограммы $\mathcal{L}_{Mel}$ [21]:

$$\mathcal{L}(\theta) = \mathcal{L}_{GAN}(\theta) + \lambda_{fm}\mathcal{L}_{FM}(\theta) + \lambda_{mel}\mathcal{L}_{Mel}(\theta) \tag{2}$$

$$\mathcal{L}(\varphi_i) = \mathcal{L}_{GAN}(\varphi_i), \quad i = 1, \ldots, k. \tag{3}$$

где $\mathcal{L}(\theta)$ обозначает функцию потерь для генератора с параметрами θ, $\mathcal{L}(\varphi_i)$ обозначает функцию потерь для i-го дискриминатора с параметрами φ_i (все дискриминаторы идентичны, за исключением разной инициализации).

Эксперименты Сравнение HiFi++ с базовыми моделями показано в Table 1. Наша модель достигает сопоставимого качества с VoiceFixer [31] и DEMUCS [10], при этом имеет значительно меньший размер. Интересно, что VoiceFixer достигает высокого субъективного качества, уступая другим моделям по объективным метрикам, особенно по SI-SDR и STOI. Действительно, VoiceFixer не использует информацию о форме волны напрямую и принимает на вход только мел-спектрограмму, таким образом, пропуская части входного сигнала и не стремясь точно восстанавливать оригинальный сигнал, что приводит к низкой производительности по классическим относительным метрикам, таким как SI-SDR, STOI и PESQ. Наша модель обеспечивает высокие относительные качественные метрики, так как явно использует исходную форму волны сигнала в качестве входных данных. В то же время наша модель учитывает спектр сигнала, что очень информативно для улучшения речи, как было показано успехом классических методов, основанных на анализе спектра. Примечательно, что мы значительно превосходим модель SEANet [44], которая обучается аналогичным состязательным образом и имеет большее количество параметров, но не учитывает спектральную информацию.

2.1.3 FFC-SE: Свертка над быстрым преобразованием Фурье для улучшения речи

Мы дополнительно улучшаем результаты HiFi++ путем разработки новых нейронных архитектур на основе оператора свертки над быстрым преобразованием Фурье (англ., Fast Fourier Convolution, FFC) [7], который мы адаптируем для задач улучшения речи. Слои FFC изначально были предложены для задач компьютерного зрения в качестве нелокального оператора, заменяющего стандартные сверточные слои в существующих нейронных сетях. Быстрое преобразование Фурье обладает глобальным рецептивным полем и было показано, что оно полезно для восстановления периодических фонов в задачах заполнения изображений [43]. Эти свойства FFC

Таблица 1: Результаты устранения шума в речи на наборе данных Voicebank-DEMAND.

Model	MOS	WV-MOS	SI-SDR	STOI	PESQ	DNSMOS	# Par (M)
Ground truth	4.60 ± 0.03	4.50	-	1.00	4.64	3.15	-
HiFi++ (ours)	$\mathbf{4.33 \pm 0.06}$	4.27	18.4	**0.95**	2.76	3.10	**1.7**
VoiceFixer	$\mathbf{4.32 \pm 0.05}$	4.14	-18.5	0.89	2.38	3.13	122.1
DEMUCS	4.22 ± 0.05	**4.37**	**18.5**	**0.95**	3.03	**3.14**	60.8
MetricGAN+	4.01 ± 0.09	3.90	8.5	0.93	**3.13**	2.95	2.7
SEANet	3.99 ± 0.09	4.19	13.5	0.92	2.36	3.05	9.2
SE-Conformer	3.39 ± 0.09	3.88	15.8	0.91	2.16	2.85	1.8
Input	3.36 ± 0.06	2.99	8.4	0.92	1.97	2.53	-

особенно полезны для предсказания комплексных спектрограмм. Действительно, гармоники спектрограмм имеют периодическую структуру, которые могут естественным образом обрабатываться с помощью сверток на быстрым преобразованием Фурье (см. Рис. 3). Кроме того, мы экспериментально наблюдаем, что большое рецептивное поле FFC полезно для восстановления фазовой информации.

Основываясь на этих выводах, были разработаны новые нейронные архитектуры для непосредственного оценивания комплексных спектрограмм в задачах улучшения речи. Предложенные модели достигают передовых результатов на наборах данных VoiceBank-DEMAND [46] и Deep Noise Suppression [12] с гораздо меньшим количеством параметров по сравнению с моделями из литературы.

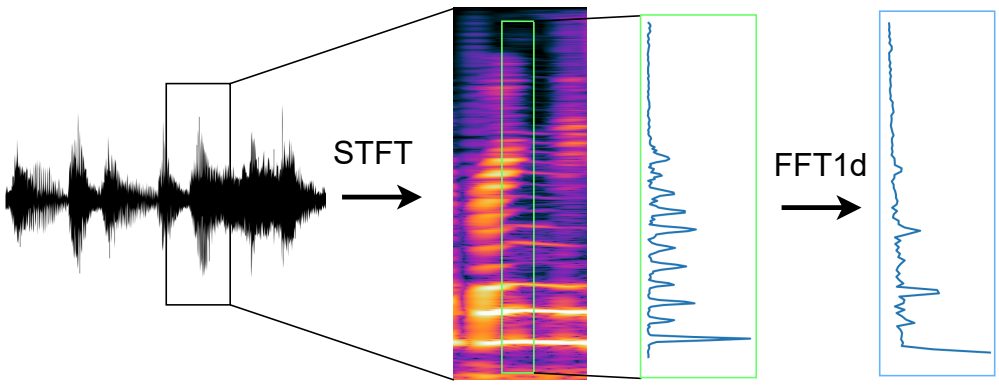

Рис. 3: Гармоники оконного преобразования Фурье образуют периодические структуры, которые могут естественным образом обрабатываться в частотной области глобальной ветвью сверток над быстрым преобразованием Фурье.

Свертки над быстрым преобразованием Фурье Свертка над быстрым преобразованием Фурье (FFC) [7] является нейронным оператором, который позволяет выполнять нелокальные вычисления и генерацию внутри нейронной сети. FFC использует поканальное быстрое преобразование Фурье [36], за которым следует точечная свертка и обратное преобразование Фурье, таким образом глобально воздействуя на входной тензор по измерениям, участвующим в преобразовании Фурье. FFC разделяет каналы на локальные и глобальные ветви. Локальная ветвь использует обычные свертки для локальных обновлений тензоров признаков, в то время как глобальная ветвь выполняет преобразование Фурье над тензором признаков и обновляет его в спектральной области, воздействуя на глобальный контекст.

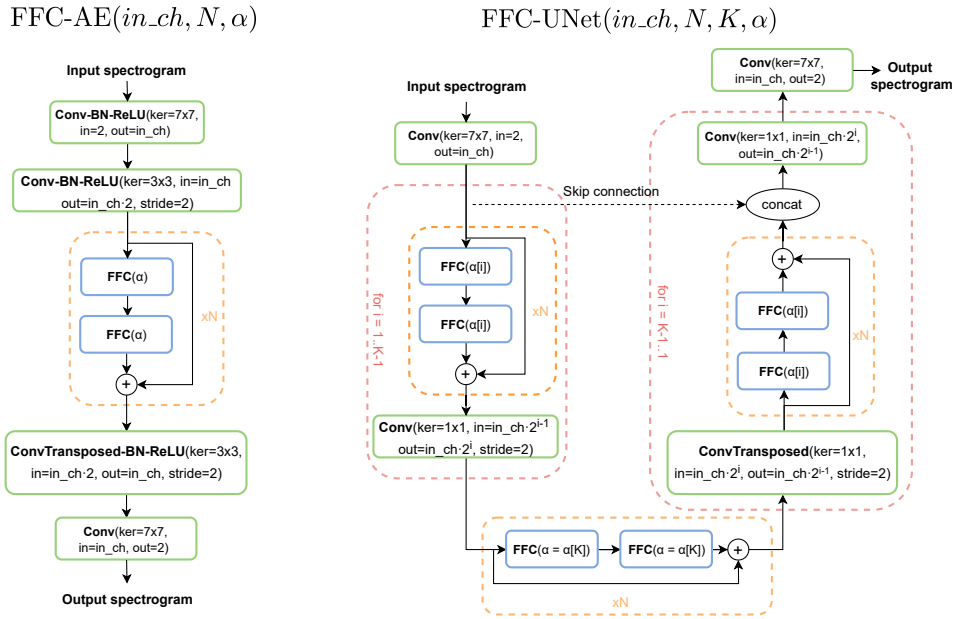

Рис. 4: Предложенные архитектуры для улучшения речи. *Слева:* автокодировщик на основе сверток над быстрым преобразованием Фурье, использующий архитектуру, предложенную в [43], для задачи улучшения речи. *Справа:* U-Net на основе сверток над быстрым преобразованием Фурье.

Архитектуры Были реализованы две нейронные сети для улучшения речи. Первая (FFC-AE) вдохновлена работой [43]. Эта архитектура состоит из сверточного кодировщика, который уменьшает размерность входного представления STFT по временной и частотной осям в два раза. За кодировщиком следует серия остаточных блоков, каждый из которых состоит из двух последовательных модулей сверток над быстрым преобразованием Фурье. Выход остаточных блоков затем увеличивается с помощью транспонированной свёртки и используется для предсказания действительных и

мнимых частей очищенной комплексно-значной спектрограммы. Архитектура показана на Figure 4 (слева). Мы называем эту модель автокодировщиком на основе сверток над быстрым преобразованием Фурье (FFC-AE). Вторая архитектура вдохновлена классической работой [38]. Мы включаем слои FFC в архитектуру U-Net, как показано на Figure 4 (справа). На каждом уровне структуры U-Net мы используем несколько остаточных блоков FFC с сверточным увеличением или уменьшением разрешения тензоров.

Таблица 2: Результаты удаления шума на датасете Voicebank-DEMAND. 3 лучших результата указаны жирным шрифтом.

Model	MOS	WV-MOS	SI-SDR	PESQ	# Params (M)
Ground Truth	4.46 ± 0.06	4.50	-	4.64	-
Input	3.44 ± 0.06	2.99	8.4	1.97	-
MetricGAN+ [15]	3.82 ± 0.06	3.90	8.5	**3.13**	2.7
ResUNet-Decouple+ [22]	3.94 ± 0.04	4.13	**18.4**	2.45	102.6
DEMUCS (non-caus.) [10]	4.06 ± 0.03	**4.37**	**18.5**	**3.03**	60.8
VoiceFixer [31]	4.10 ± 0.03	4.14	-18.5	2.38	122.1
HiFi++	4.15 ± 0.07	4.27	**18.4**	2.76	**1.7**
FFC-AE-V0 (ours)	$\mathbf{4.24 \pm 0.09}$	4.34	17.9	2.88	**0.42**
FFC-AE-V1 (ours)	$\mathbf{4.33 \pm 0.03}$	**4.37**	17.5	2.96	**1.7**
FFC-UNet (ours)	$\mathbf{4.28 \pm 0.03}$	**4.38**	18.1	**2.99**	7.7
FFC-AE-V1 (abl.)	3.98 ± 0.07	4.05	16.7	2.68	2.9
vanilla UNet	4.10 ± 0.07	4.11	17.2	2.73	20.7

Эксперименты Было проведено сравнение качества предлагаемых моделей с несколькими базовыми моделями из литературы. Как видно из Таблицы 2, на Voicebank-DEMAND наши модели значительно превосходят все базовые модели по MOS и дают конкурентные результаты по объективным метрикам. В целом, мы наблюдаем, что нейронные архитектуры, построенные на основе сверток над быстрым преобразованием Фурье, значительно превосходят архитектуры на основе обычной свертки с точки зрения качества улучшения речи, оценки фазы и эффективности использования параметров. Предлагаемые архитектуры обеспечивают передовые результаты на бенчмарках шумоподавления речи, будучи значительно меньше по размеру, чем базовые модели.

2.1.4 FINALLY: быстрое и универсальное улучшение речи с качеством студийной записи

Ранее обсуждаемые подходы демонстрируют впечатляющие результаты на записях речи, загрязненных аддитивным шумом. Однако на практике реальные записи часто загрязнены несколькими искажениями одновремен-

но. Мы обнаружили, что модели, обученные для устранения только аддитивного шума, плохо обобщаются на реальные данные. Поэтому в данном разделе мы обобщаем HiFi++ на случай нескольких деградаций одновременно и демонстрируем, что эта модель обеспечивают быстрое и высококачественное универсальное улучшение речи, то есть одна модель может использоваться для устранения нескольких видов деградаций одновременно. Наша модель превосходит как диффузионные модели, так и предыдущие модели на основе GAN, достигая беспрецедентного уровня качества как на смоделированных, так и на реальных данных.

Для достижения этого мы исследуем различные модели создания векторных признаков в качестве основы для функции потерь для генерации речи и предлагаем критерии для выбора модели на основе структуры ей пространства ее векторных признаков. Эти критерии подтверждаются эмпирическими результатами на задаче нейронного вокодинга, указывающими на то, что сверточные признаки нейронной сети WavLM хорошо подходят для функции потерь при улучшении речи. Мы также разрабатываем новую модель для универсального улучшения речи, которая интегрирует предложенную функцию потерь с тренировкой MS-STFT дискриминатора и улучшает архитектуру генератора HiFi++, объединяя ее с предобученным энкодером WavLM [6].

Архитектура Мы вводим два изменения в архитектуру генератора HiFi++ (см. Figure 5). Во-первых, мы модифицируем генератор, добавляя выход модели WavLM-large (последнее скрытое состояние трансформера) в качестве дополнительного входа для Upsampler. Предыдущие работы [18, 4] показали полезность признаков самообучения (англ. self-supervised learning, SSL) для задач улучшения речи, и мы получаем схожий результат, наблюдая значительное повышение качества при использовании признаков SSL. Во-вторых, мы добавляем Upsample WaveUNet в конце генератора. Это позволяет модели выдавать сигнал с частотой 48 кГц, принимая сигнал с частотой 16 кГц на входе.

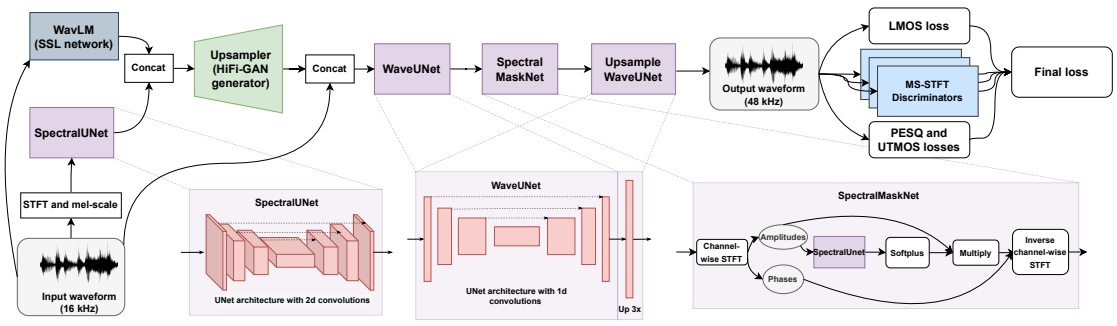

Рис. 5: Архитектура модели FINALLY.

Обучение Во время обучения, перед смешиванием с шумом, мы сворачиваем речевой сигнал с случайно выбранным импульсным откликом микрофона и применяем другие цифровые искажения. Мы обучаем модель в три этапа. Первые два этапа сосредоточены на восстановлении исходного содержания речи, а заключительный этап направлен на улучшение эстетического восприятия речи. Функции потерь, которые мы используем, могут быть записаны следующим образом:

$$\mathcal{L}_{\text{LMOS}}(\theta) = \mathop{\mathbb{E}}_{x,y\sim p(x,y)} \left[100 \cdot \|\phi(y) - \phi(g_\theta(x))\|_2^2 + \||\text{STFT}(y)| - |\text{STFT}(g_\theta(x))|\|_1\right], \quad (4)$$

$$\mathcal{L}_{\text{gen}}(\theta) = \overbrace{\underbrace{\lambda_{\text{LMOS}} \cdot \mathcal{L}_{\text{LMOS}}(\theta)}_{\text{1st stage (16 kHz)}} + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN-gen}}(\theta) + \lambda_{\text{FM}} \cdot \mathcal{L}_{\text{FM}}(\theta)}^{\text{2nd stage (16 kHz)}} + \lambda_{\text{HF}} \cdot \mathcal{L}_{\text{HF}}(\theta), \quad (5)$$

$$\mathcal{L}_{\text{disc}}(\varphi_i) = \mathcal{L}_{\text{GAN-disc}}(\varphi_i), \quad i = 1, \ldots, k. \quad (6)$$

Здесь, ϕ обозначает функцию получения признаков WavLM-conv, $g_\theta(x)$ обозначает нейронную сеть генератора с параметрами θ, $\mathcal{L}_{\text{GAN-gen}}(\theta)$ обозначает функцию потерь генератора LS-GAN [33], $\mathcal{L}(\theta)$ обозначает комбинированную функцию потерь генератора, $\mathcal{L}_{\text{GAN-disc}}(\varphi_i)$ обозначает функцию потерь дискриминатора LS-GAN [33] для i-го дискриминатора с параметрами φ_i, $\mathcal{L}_{\text{FM}}$ обозначает функцию потерь по сопоставлению признаков [24, 11], $\mathcal{L}_{\text{HF}}$ обозначает функцию потерь по восприятию людей, а λ_* обозначает соответствующие веса функций потерь.

Таблица 3: Сравнение с методами из литературы.

VoxCeleb (HiFi-GAN-2 validation set, real data)						
Model	MOS	DNSMOS	UTMOS	WV-MOS	-	RTF
Input	3.46 ± 0.07	2.72 ± 0.11	2.76 ± 0.13	2.90 ± 0.16	-	-
VoiceFixer	3.41 ± 0.07	3.08 ± 0.06	2.60 ± 0.09	2.79 ± 0.09	-	0.02
DEMUCS	3.79 ± 0.07	3.27 ± 0.04	3.51 ± 0.08	3.72 ± 0.08	-	0.08
STORM	3.75 ± 0.06	3.17 ± 0.04	3.29 ± 0.08	3.54 ± 0.09	-	1.05
BBED	3.97 ± 0.06	3.23 ± 0.04	3.30 ± 0.10	3.47 ± 0.08	-	0.43
HiFi-GAN-2	4.47 ± 0.05	**3.32 ± 0.03**	3.67 ± 0.09	**3.96 ± 0.06**	-	0.50
Ours	**4.63 ± 0.04**	**3.31 ± 0.04**	**4.05 ± 0.07**	**3.98 ± 0.06**	-	**0.03**
UNIVERSE validation set (simulated data)						
Model	MOS	DNSMOS	UTMOS	WV-MOS	PhER	RTF
Input	2.87 ± 0.05	2.25 ± 0.19	2.27 ± 0.28	1.72 ± 0.61	0.31 ± 0.05	-
Ground Truth	4.39 ± 0.05	3.33 ± 0.04	4.26 ± 0.06	4.28 ± 0.06	0	-
UNIVERSE	4.10 ± 0.07	**3.23 ± 0.07**	3.89 ± 0.15	3.85 ± 0.12	0.20 ± 0.04	0.5
Ours (16 kHz)	3.99 ± 0.07	**3.24 ± 0.05**	**4.21 ± 0.08**	**4.43 ± 0.07**	**0.14 ± 0.03**	**0.03**
Ours	**4.23 ± 0.07**	**3.25 ± 0.05**	**4.21 ± 0.10**	**4.43 ± 0.08**	**0.14 ± 0.03**	**0.03**

Эксперименты Мы рассматриваем модели диффузии BBED [26], STORM [28] и UNIVERSE [40], а также регрессионные модели Voicefixer и DEMUCS в качестве базовых моделей из прошлых работ. В дополнение к ним мы рассматриваем ближайшего конкурента, HiFi-GAN-2, в качестве базовой модели на основе GAN. Данные для сравнения с HiFi-GAN-2 и UNIVERSE были взяты с их демонстрационных страниц, так как авторы не выкладывали исходный код. Мы проводим сравнения с BBED, STORM, Voicefixer, DEMUCS и HiFi-GAN-2 на реальных образцах VoxCeleb1, а сравнение с UNIVERSE на смоделированных данных, предоставленных авторами этой работы. Мы также предоставляем результаты для наших предсказаний, пересэмплированных к 16 кГц, в дополнение к базовым предсказаниям в 48 кГц, так как модель UNIVERSE выдает только треки с частотой 16 кГц. Сравнение представлено в Таблице 3.

Таким образом, интегрируя новую функцию потерь на основе WavLM с состязательным обучением на основе MS-STFT дискриминаторов и улучшая архитектуру HiFi++ с помощью энкодера WavLM, мы разработали новую модель улучшения речи, FINALLY, которая достигает передовых результатов, восстанавливая четкую и высококачественную речь на частоте 48 кГц.

2.2 Итеративная авторегрессия для улучшения речи в потоковом режиме

Потоковая генерация по своей природе следует последовательной структуре предсказаний, что хорошо подходит для авторегрессии. Традиционный подход к обучению авторегрессионных моделей заключается в использовании метода "обучение с принуждением учителя"(англ. teacher forcing) [47], при котором модель во время обучения использует предыдущие значения из истинных данных (англ. groundtruth) для прогнозирования следующих. На этапе предсказаний модель использует предсказанные значения для авторегрессионного обуславливания, поскольку истинные значения недоступны. Обучение с принуждением учителя является эффективным способом обучения, поскольку его можно параллелизовать для сетей на основе сверток. Однако его основным недостатком является несоответствие между стадиями обучения и предсказаний (вывода), что может привести к значительному снижению качества в тестовой фазе. Было замечено, что модели улучшения речи на основе авторегрессии очень уязвимы к несоответствию между обучением и выводом. Большинство подходов в литературе решают эту проблему, используя собственные предсказания модели для авторегресионного обуславливания во время обучения [2]. Однако эти методы обычно используются для рекуррентных сетей в задачах низкой раз-

мерности, поскольку их применение к сетям на основе сверток существенно замедляет обучение при высокой размерности, что мешает практическому применению.

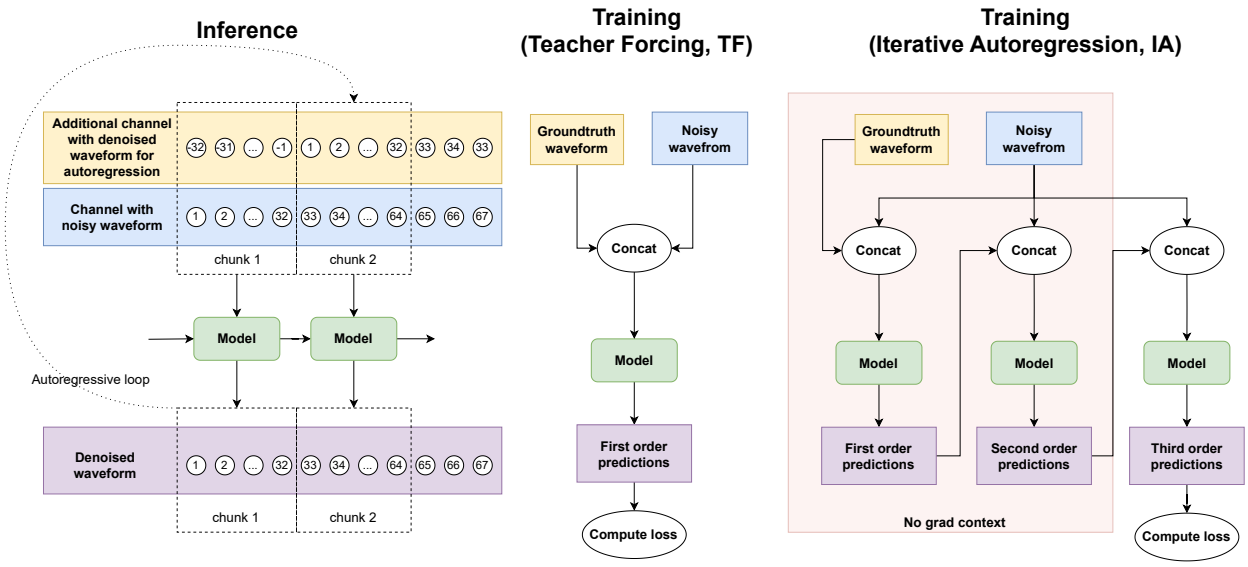

Рис. 6: *Слева:* иллюстрация авторегрессионного обуславливания для модели с алгоритмической задержкой 32 временных шага (2 мс при частоте дискретизации 16 кГц). Предсказанный сигнал из блока 1 повторно используются при составлении прогнозов для блока 2 во время предсказаний. *Посередине:* Иллюстрация обучения с принуждением учителя. *Справа:* Этап 3 процесса итеративной авторегрессии предложенной в данной работе. Модель использует свои собственные предсказания для получения прогнозов более высоких порядков. Мы сдвигаем истинную форму волны и прогнозы перед формированием канала с авторегрессионным условием, чтобы избежать утечки информации о будущем.

Итеративная авторегрессия Мы представляем простой, но весьма эффективный алгоритм для обучения авторегрессионных моделей, который значительно снижает несоответствие между этапами обучения и предсказаний (см. Рис. 6). Наш подход основан на итеративной замене авторегрессионного обуславливания с истинных данных на прогнозы модели в режиме обучения с принуждением учителя. А именно, мы делим весь процесс обучения на N этапов, начиная со стандартного обучения с принуждением учителя на начальном этапе. На втором этапе прямой проход модели включает два шага: на первом шаге модель делает прогноз, основанный на истинных данных, на втором шаге - делает прогноз, основанный на предсказаниях с первого шага. Аналогично, на n-м этапе прямой проход состоит из n шагов, и на каждом шаге модель основывается на предска-

заниях с предыдущего шага. Мы называем этот алгоритм "итеративной авторегрессией"(IA) и демонстрируем, что он помогает уменьшить несоответствие между этапами обучения и предсказаний, вызванное обучением с принуждением учителя. Более того, мы показываем, что авторегрессионное условие предлагает значительные преимущества по сравнению с неавторегрессионными базовыми моделями при различных функциях потерь в обучении и архитектурах нейронных сетей. Примечательно, что предложенный нами алгоритм IA универсален и может быть потенциально применен к обучению авторегрессионных моделей за пределами области улучшения речи.

Таблица 4: Авторегрессионное обучение улучшает качество удаления шума в разных сценариях. Все авторегрессионные модели (AR) обучены с использованием предложенного метода итеративной авторегрессии, если не обозначено обучение с использованием принуждения учителя (TF).

Experiment	**UTMOS**	**DNSMOS**	**SISDR**	**CMOS**
Base configuration				
w/o AR	3.53	2.97	17.0	-
w/ AR (TF)	3.38	2.92	12.6	$-0.5^{\pm 0.08}$
w/ AR	**3.61**	**3.03**	**18.4**	$0.1^{\pm 0.05}$
Different losses				
w/o AR (adv.)	3.68	3.02	15.2	-
w/ AR (adv.)	**3.74**	**3.04**	**15.3**	$0.12^{\pm 0.04}$
w/o AR (si-snr)	3.51	2.95	17.0	-
w/ AR (si-snr)	**3.57**	**2.96**	**17.8**	$0.13^{\pm 0.05}$
DNS dataset				
w/o AR	2.42	2.98	14.5	-
w/ AR	**2.47**	**3.03**	**14.6**	$0.1^{\pm 0.05}$
ConvTasNet architecture				
w/o AR	3.08	2.86	15.3	-
w/ AR	**3.33**	**2.99**	**15.8**	$0.52^{\pm 0.06}$
Different latencies				
w/o AR (2 ms)	3.47	2.94	17.1	-
w/ AR (2 ms)	**3.55**	**2.98**	**18.3**	$0.04^{\pm 0.04}$
w/o AR (4 ms)	3.5	2.96	17.2	-
w/ AR (4 ms)	**3.59**	**3.02**	**18.6**	$0.16^{\pm 0.05}$
w/o AR (16 ms)	3.57	2.99	17.3	-
w/ AR (16 ms)	**3.64**	**3.02**	**18.6**	$0.09^{\pm 0.04}$

Эксперименты Во всех наших экспериментах мы рассматриваем аддитивный шум как искажение, которое необходимо удалить из записей речи.

Как показано в Таблице 4, мы проводим ряд экспериментов, чтобы проверить эффективность итеративной авторегрессии в различных условиях обучения.

Для каждого экспериментального условия мы обучаем базовую модель без авторегрессионного условия (w/o AR) и модель с авторегрессионным условием (w/ AR). Условия обучения и модели идентичны, за исключением того, что AR модели обучаются с использованием итеративной авторегрессии, если не указано иное.

Эксперименты можно разделить на 5 групп в зависимости от применяемого сценария обучения. В каждом сценарии обучения мы изменяем только одно условие обучения (набор данных/архитектура модели/функция потерь/задержка), оставляя остальные параметры такими же, как в базовой конфигурации, описанной ниже.

Предложенный метод итеративного авторегрессионного обучения позволяет улучшить качество моделей улучшения речи в потоковом режиме во всех изученных сценариях. Более того, он значительно превосходит традиционный метод обучения с принуждением, который не дает никаких улучшений по сравнению с неавторегрессионной базовой моделью из-за высокого несоответствия между обучением и предсказаниями. Мы считаем, что представленная техника обеспечивает практическую альтернативу обучению с принуждением учителя и представляет важный шаг к улучшению потоковых моделей с помощью авторегрессии.

2.3 Улучшение речи без учителя с использованием безусловной модели диффузии

В последние годы диффузионные модели [41, 17, 20] привлекли внимание исследователей благодаря своей способности эффективно моделировать сложные высокоразмерные распределения. Диффузионные модели моделируют априорное распределение данных, посредством выучивания градиента логарифма плотности зашумленных данных. Полученное априорное распределение может быть полезным для решения обратных задач, где целью является восстановление входного сигнала y из измерений x, которые обычно связаны через некоторый дифференцируемый оператор A, т.е. $x = A(y) + n$, где n - это некоторый шум. В этой части работы мы представляем UnDiff, диффузионную вероятностную модель, разработанную для решения различных обратных задач в обработке речи, включая инверсию деградации и нейронное вокодирование.

Ключевой особенностью UnDiff является ее способность обучаться в безусловном режиме для генерации волновых форм речи и затем адаптироваться к обратной задаче без какого-либо дополнительного обучения. Это

отличает Undiff от существующих подходов, которые используют условные диффузионные модели для восстановления и генерации речи или разрабатывают специфические обучающие схемы для конкретных задач [40, 37, 39].

Решение обратных задач с помощью диффузионных моделей Обратные задачи решают задачу восстановления объекта $\mathbf{y}$ по его частичному наблюдению $\mathbf{x}$ и модели деградации $p(\mathbf{x}|\mathbf{y})$. Чтобы использовать обратное СДУ для сэмплирования из условного распределения $p(\mathbf{y}|\mathbf{x})$, необходимо знать скор-функцию условного распределения $\nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t|\mathbf{x})$.

Один из способов оценить $\nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t|\mathbf{x})$ - применить импутацию (согласованность данных) [41, 34, 8]. Идея этого метода заключается в явном изменении скор-функции таким образом, чтобы некоторые части оценки сэмпла без шума $\hat{\mathbf{y}}_0 = \frac{1}{\sqrt{\bar{\alpha}(t)}}(\mathbf{y}_t - (1 - \bar{\alpha}(t))s_\theta(\mathbf{y}_t, t))$ были заполнены наблюдениями $\mathbf{x}$.

Другой способ формализовать поиск $\mathbf{y}$ - использование формулы Байеса:

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})/p(\mathbf{x}), \tag{7}$$

следовательно,

$$\nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t|\mathbf{x}) = \nabla_{\mathbf{y}_t} \log p_t(\mathbf{x}|\mathbf{y}_t) + \nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t), \tag{8}$$

$\nabla_{\mathbf{y}_t} \log p_t(\mathbf{x}|\mathbf{y}_t)$ обычно невычислимо. Однако, Чхун и др. [9] показали, что можно приблизить $\nabla_{\mathbf{y}_t} \log p(\mathbf{x}|\mathbf{y}_t) \approx \nabla_{\mathbf{y}_t} \log p(\mathbf{x}|\hat{\mathbf{y}}_0)$, где $\nabla_{\mathbf{y}_t} \log p(\mathbf{x}|\hat{\mathbf{y}}_0)$ можно вычислить, используя модель деградации. Учитывая оператор наблюдения A и предполагая гауссовское распределение наблюдений, конечное приближение становится:

$$\nabla_{\mathbf{y}_t} \log p_t(\mathbf{x}|\mathbf{y}_t) \approx -\xi(t)\nabla_{\mathbf{y}_t}\|\mathbf{x} - A(\hat{\mathbf{y}}_0)\|_2^2 \tag{9}$$

где $\xi(t)$ является весовым коэффициентом, который мы устанавливаем обратно пропорционально норме градиента, аналогично [34]. Как и [34], мы называем этот метод реконструкционным руководством.

Расширение полосы пропускания Расширение полосы пропускания частот [23, 1] (также известное как аудио супер-разрешение) можно рассматривать как реалистичное восстановление высоких частот волновой формы. Оператор наблюдения является фильтром нижних частот $\mathbf{x} = A(\mathbf{y}) = \mathrm{LPF}(\mathbf{y})$. Таким образом, обуславливание в данном случае соответствует замене сгенерированных нижних частот наблюдаемыми нижними

частотами $\mathbf{x}$ на каждом шаге. Более формально это соответствует изменению скор-функции во время решения обратного СДУ следующим образом:

$$\tilde{s}_\theta(\mathbf{y}_t, t) = \frac{1}{1-\bar{\alpha}(t)}(\sqrt{\bar{\alpha}(t)}\tilde{\hat{\mathbf{y}}}_0 - \mathbf{y}_t), \tag{10}$$

где $\tilde{\hat{\mathbf{y}}}_0 = \hat{\mathbf{y}}_0 - \mathrm{LPF}(\hat{\mathbf{y}}_0) + \mathbf{x}$ является дополненной оценкой $\mathbf{y}_0$, а $\hat{\mathbf{y}}_0 = \frac{1}{\sqrt{\bar{\alpha}(t)}}(\mathbf{y}_t + (1-\bar{\alpha}(t))s_\theta(\mathbf{y}_t, t))$ является оценкой $\mathbf{y}_0$ с исходной скор-функцией. В наших экспериментах по расширению полосы пропускания мы используем записи с частотой дискретизации 16 кГц в качестве целевых и рассматриваем две полосы пропускания частот для входных данных: 2 кГц и 4 кГц. Мы искусственно ухудшаем сигнал до желаемой полосы пропускания частот (2 кГц или 4 кГц) с использованием полифазной фильтрации. Результаты и сравнение с другими методами для 2 кГц представлены в Таблице 5.

Таблица 5: Результаты расширения полосы пропускания (BWE 2 кГц) на VCTK.

Model	Supervised	WV-MOS	LSD	MOS
Ground Truth	-	4.17	0	4.09 ± 0.09
HiFi++	✓	**4.05**	1.09	$\mathbf{3.93 \pm 0.10}$
Voicefixer [31]	✓	3.67	1.08	3.64 ± 0.10
TFiLM [3]	✓	2.83	1.01	2.71 ± 0.10
UnDiff (Diffwave)	×	3.48	**0.96**	3.59 ± 0.11
UnDiff (FFC-AE)	×	3.59	1.13	3.50 ± 0.11

Деклиппинг Мы рассматриваем обрезание (клиппинг) как обратную задачу с функцией наблюдения, определенной как $A = \mathrm{clip}(\mathbf{y}) = \frac{1}{2}(|y+c| - |y-c|)$ и применяем стратегию реконструкционного руководства. Мы сравниваем наши модели с популярными методами деклиппинга A-SPADE [49] и S-SPADE [50], а также с общей системой восстановления речи Voicefixer [31] на обрезанных аудиозаписях с входным SDR, равным 3 дБ (см. Таблицу 6).

Результаты показывают, что, несмотря на то, что UnDiff никогда не обучалась явно для решения каких-либо из рассматриваемых задач, она демонстрирует сопоставимые с (хотя и уступающие) базовыми моделями результаты для расширения полосы пропускания и деклиппинга. В целом, результаты подчеркивают потенциал моделей безусловной диффузии как универсальных инструментов для восстановления речи без учителя.

Таблица 6: Результаты деклиппинга (входной SNR = 3 дБ) на VCTK.

Model	Supervised	WV-MOS	SI-SNR	MOS
Ground Truth	-	3.91	-	3.84 ± 0.11
A-SPADE [49]	×	2.63	8.48	2.67 ± 0.11
S-SPADE [50]	×	2.69	8.50	2.55 ± 0.11
Voicefixer [31]	✓	2.79	-22.58	2.98 ± 0.12
UnDiff (Diffwave)	×	**3.62**	**10.57**	$\mathbf{3.59 \pm 0.12}$
UnDiff (FFC-AE)	×	3.01	7.35	3.06 ± 0.12
Input	-	2.30	3.82	2.19 ± 0.09

3 Выводы

Основные выводы, сделанные на основе результатов этой работы, следующие:

1. Мультидоменные архитектуры генераторов обеспечивают лучшее соотношение между качеством и вычислительной сложностью моделей расширения полосы пропускания (BWE) и улучшения речи (SE). В частности, полезно усиливать модели обработки аудио модулями, которые выполняют коррекцию сигнала как в временной, так и в спектральной области, для более эффективного использования параметров и снижения вычислительной сложности, как показано в исследовании HiFi++.

2. Блоки свертки над быстрым преобразованием Фурье (англ. Fast Fourier Convolution) предоставляют эффективный архитектурный выбор для разработки модулей обработки спектра. Глобальное рецептивное поле этого нейронного слоя позволяет эффективно оценивать фазу, снижая потребление памяти для весов нейронной сети.

3. Теоретический анализ показывает, что состязательное обучение на основе LS-GAN может использоваться для неявной регрессии на моду распределения, что естественно согласуется с практическими целями задачи улучшения речи. Экспериментальная реализация обучения на основе LS-GAN подтверждает этот анализ и показывает, что модели на основе GAN способны достигать быстрого и качественного улучшения речи, превосходя другие типы генеративных моделей при меньших ресурсах.

4. Авторегрессионное обуславливание способно улучшить модели улучшения речи в потоковом режиме, используя информацию о прошлых

предсказаниях во время вывода. Однако применение стандартной техники обучения авторегрессионных моделей, которая называется обучением с принуждением учителя, приводит к высокому уровню несоответствия между обучением и предсказаниями и, следовательно, к низкому качеству улучшения речи. Разработанная техника итеративной авторегрессии предоставляет практическую альтернативу обучению с принуждением учителя и позволяет эффективно обучать авторегрессионные модели улучшения речи.

5. Улучшение речи без учителя представляет значительную проблему из-за неизвестной модели деградации во время обучения. Эту проблему можно решить с помощью модели безусловной диффузии. Модель безусловной диффузии может быть использована для моделирования априорного распределения речевых сигналов во время обучения, а затем адаптироваться к конкретной модели деградации во время предсказаний. К сожалению, наше исследование выявляет значительные сложности, возникающие при таком подходе, и итоговые модели, как правило, работают значительно хуже, чем их аналоги, обученные с учителем.

Список литературы

[1] Pavel Andreev и др. «Hifi++: a unified framework for bandwidth extension and speech enhancement». В: *arXiv preprint arXiv:2203.13086* (2022).

[2] Samy Bengio и др. «Scheduled sampling for sequence prediction with recurrent neural networks». В: *Advances in neural information processing systems* 28 (2015).

[3] Sawyer Birnbaum и др. «Temporal film: Capturing long-range sequence dependencies with feature-wise modulations». В: *arXiv preprint arXiv:1909.06628* (2019).

[4] Jaeuk Byun и др. «An Empirical Study on Speech Restoration Guided by Self-Supervised Speech Representation». В: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, с. 1—5.

[5] Jun Chen и др. «Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement». В: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, с. 7857—7861.

[6] Sanyuan Chen и др. «Wavlm: Large-scale self-supervised pre-training for full stack speech processing». В: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), с. 1505—1518.

[7] Lu Chi, Borui Jiang и Yadong Mu. «Fast fourier convolution». В: *Advances in Neural Information Processing Systems* 33 (2020), с. 4479—4488.

[8] Jooyoung Choi и др. «ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models». В: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, с. 14367—14376.

[9] Hyungjin Chung и др. «Diffusion posterior sampling for general noisy inverse problems». В: *arXiv preprint arXiv:2209.14687* (2022).

[10] Alexandre Defossez, Gabriel Synnaeve и Yossi Adi. «Real Time Speech Enhancement in the Waveform Domain». В: *Interspeech*. 2020.

[11] Alexandre Défossez и др. «High Fidelity Neural Audio Compression». В: *Transactions on Machine Learning Research* (2023).

[12] Harishchandra Dubey и др. «Icassp 2022 deep noise suppression challenge». В: *arXiv preprint arXiv:2202.13288* (2022).

[13] Yariv Ephraim. «Statistical-model-based speech enhancement systems». В: *Proceedings of the IEEE* 80.10 (1992), с. 1526—1555.

[14] Yariv Ephraim и David Malah. «Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator». В: *IEEE Transactions on acoustics, speech, and signal processing* 32.6 (1984), с. 1109—1121.

[15] Szu-Wei Fu и др. «MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement». В: *arXiv preprint arXiv:2104.03538* (2021).

[16] Xiang Hao и др. «FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement». В: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, с. 6633—6637.

[17] Jonathan Ho, Ajay Jain и Pieter Abbeel. «Denoising diffusion probabilistic models». В: *Advances in Neural Information Processing Systems* 33 (2020), с. 6840—6851.

[18] Kuo-Hsuan Hung и др. «Boosting self-supervised embeddings for speech enhancement». В: *arXiv preprint arXiv:2204.03339* (2022).

[19] Umut Isik и др. «Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss». В: *arXiv preprint arXiv:2008.04470* (2020).

[20] Tero Karras и др. «Elucidating the Design Space of Diffusion-Based Generative Models». В: *Advances in Neural Information Processing Systems*.

[21] Jungil Kong, Jaehyeon Kim и Jaekyoung Bae. «Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis». В: *arXiv preprint arXiv:2010.05646* (2020).

[22] Qiuqiang Kong и др. «Decoupling magnitude and phase estimation with deep resunet for music source separation». В: *arXiv preprint arXiv:2109.05418* (2021).

[23] Volodymyr Kuleshov, S Zayd Enam и Stefano Ermon. «Audio super resolution using neural networks». В: *arXiv preprint arXiv:1708.00853* (2017).

[24] Kundan Kumar и др. «Melgan: Generative adversarial networks for conditional waveform synthesis». В: *arXiv preprint arXiv:1910.06711* (2019).

[25] Anders Boesen Lindbo Larsen и др. «Autoencoding beyond pixels using a learned similarity metric». В: *International conference on machine learning*. PMLR. 2016, с. 1558—1566.

[26] Bunlong Lay и др. «Reducing the Prior Mismatch of Stochastic Differential Equations for Diffusion-based Speech Enhancement». В: *arXiv preprint arXiv:2302.14748* (2023).

[27] Christian Ledig и др. «Photo-realistic single image super-resolution using a generative adversarial network». В: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, с. 4681—4690.

[28] Jean-Marie Lemercier и др. «StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation». В: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[29] Cheuk Ting Li и Farzan Farnia. «Mode-seeking divergences: theory and applications to GANs». В: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, с. 8321—8350.

[30] Hsin-Yi Lin и др. «Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport». В: *Advances in Neural Information Processing Systems* 34 (2021), с. 19935—19946.

[31] Haohe Liu и др. «VoiceFixer: Toward General Speech Restoration with Neural Vocoder». В: *arXiv preprint arXiv:2109.13731* (2021).

[32] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.

[33] Xudong Mao и др. «Least squares generative adversarial networks». В: *Proceedings of the IEEE international conference on computer vision*. 2017, с. 2794—2802.

[34] Eloi Moliner, Jaakko Lehtinen и Vesa Välimäki. «Solving Audio Inverse Problems with a Diffusion Model». В: *arXiv preprint arXiv:2210.15228* (2022).

[35] Eloi Moliner, Jaakko Lehtinen и Vesa Välimäki. «Solving audio inverse problems with a diffusion model». В: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, с. 1—5.

[36] Henri J Nussbaumer. «The fast Fourier transform». В: *Fast Fourier Transform and Convolution Algorithms*. Springer, 1981, с. 80—111.

[37] Julius Richter и др. «Speech enhancement and dereverberation with diffusion-based generative models». В: *arXiv preprint arXiv:2208.05830* (2022).

[38] Olaf Ronneberger, Philipp Fischer и Thomas Brox. «U-net: Convolutional networks for biomedical image segmentation». В: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, с. 234—241.

[39] Robin Scheibler и др. «Diffusion-based Generative Speech Source Separation». В: *arXiv preprint arXiv:2210.17327* (2022).

[40] Joan Serrà и др. «Universal speech enhancement with score-based diffusion». В: *arXiv preprint arXiv:2206.03065* (2022).

[41] Yang Song и др. «Score-Based Generative Modeling through Stochastic Differential Equations». В: *International Conference on Learning Representations*.

[42] Daniel Stoller, Sebastian Ewert и Simon Dixon. «Wave-u-net: A multi-scale neural network for end-to-end audio source separation». В: *arXiv preprint arXiv:1806.03185* (2018).

[43] Roman Suvorov и др. «Resolution-robust Large Mask Inpainting with Fourier Convolutions». В: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, с. 2149—2159.

[44] Marco Tagliasacchi и др. «SEANet: A multi-modal speech enhancement network». В: *arXiv preprint arXiv:2009.02095* (2020).

[45] Zehai Tu и др. «A Two-Stage End-to-End System for Speech-in-Noise Hearing Aid Processing». В: *Proc. Clarity* (2021), с. 3—5.

[46] Cassia Valentini-Botinhao и др. «Noisy speech database for training speech enhancement algorithms and tts models». В: (2017).

[47] Ronald J Williams и David Zipser. «A learning algorithm for continually running fully recurrent neural networks». В: *Neural computation* 1.2 (1989), с. 270—280.

[48] Yong Xu и др. «A regression approach to speech enhancement based on deep neural networks». В: *IEEE/ACM transactions on audio, speech, and language processing* 23.1 (2014), с. 7—19.

[49] Pavel Zaviska и Pavel Rajmic. «Analysis Social Sparsity Audio Declipper». В: *arXiv preprint arXiv:2205.10215* (2022).

[50] Pavel Zaviska и др. «A proper version of synthesis-based sparse audio declipper». В: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, с. 591—595.

[51] K Zmolikova и JH Cernock. «BUT System for the First Clarity Enhancement Challenge». В: *Proceedings of Clarity* (2021), с. 1—3.