Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский университет ИТМО»

Факультет информационных технологий и программирования
Кафедра Речевых информационных систем

Отчёт по курсовой работе

Научно исследовательская работа по теме:

«Разработка безэталонного метода объективной оценки индекса разборчивости речи»

Выполнил студент гр. M4121:
Давыдов Д. А. _____
Руководитель:
Столбов М. Б. _____

г. Санкт-Петербург
27 июня 2025 г.

Содержание

Введение

Разборчивость речи является критически важным параметром для профессиональных дикторов, определяющим эффективность передачи информации аудитории. В настоящее время для оценки разборчивости речи применяются как субъективные, так и объективные методы.

Субъективные методы, такие как экспертные оценки или тестирование на группах слушателей, обладают высокой точностью, но требуют значительных временных и организационных затрат. Объективные методы, включая алгоритмические и инструментальные подходы, позволяют автоматизировать процесс оценки, однако часто зависят от эталонных данных и не всегда учитывают специфику профессиональной дикторской речи.

Учитывая недостатки существующих подходов, актуальной задачей является разработка специализированного метода оценки разборчивости речи, адаптированного именно для дикторов. **Целью данного исследования** является создание **безэталонного метода оценки разборчивости речи**, который позволит объективно и эффективно анализировать качество произношения дикторов в различных условиях без необходимости использования эталонных записей. Такой подход упростит процесс оценки и сделает его более доступным для применения в профессиональной среде.

Задачи исследования:

1. Анализ существующих субъективных и объективных методов оценки разборчивости речи (РР) и выявление их ограничений применительно к дикторской речи.

2. Разработка архитектуры нейронной сети, способной оценивать РР без использования эталонных речевых образцов.

3. Сбор и создание корпуса аудио данных для обуения нейронной сети.

4. Экспериментальная проверка точности и устойчивости предложенного метода на корпусе дикторских записей в различных акустических условиях.

Объект исследования – разборчивость речи профессиональных дикторов как ключевая характеристика качества устной коммуникации. В работе рассматриваются акустические и лингвистические особенности дикторской речи, влияющие на её восприятие, а также технические аспекты автоматизированной оценки этого параметра.

Предмет исследования – методы и алгоритмы безэталонной оценки разборчивости речи на основе нейросетевых технологий. Особое внимание уделяется возможностям глубокого обучения для выявления и количественного измерения параметров РР без сравнения с эталонными образцами.

Исследование сосредоточено на разработке нейросетевого подхода к оценке PP, не требующего эталонных записей для анализа качества дикторской речи. Основной акцент делается на создании системы, способной адаптироваться к различным стилям профессионального произношения и условиям записи.

1 Теоретические основы разборчивости речи

Оценка разборчивости речи является ключевой задачей во многих областях, включая телекоммуникации, аудио-криминалистику и разработку слуховых аппаратов. Понимание того, насколько хорошо слушатель может разобрать произнесенные слова конкретного диктора в различных акустических условиях, позволяет не только оценивать качество систем связи, но и разрабатывать новые методы обработки сигналов, направленные на улучшение восприятия речи. В реальных условиях речевой сигнал часто подвергается искажениям из-за фонового шума и реверберации, что делает задачу оценки разборчивости особенно актуальной. В данном разделе будут рассмотрены фундаментальные концепции разборчивости речи, факторы, оказывающие на нее влияние, а также существующие методы ее объективной и субъективной оценки.

1.1 Понятие разборчивости речи

Разборчивость речи определяется как доля речевого сигнала, содержание которого может быть правильно распознано слушателем [10]. Этот показатель является важным количественным параметром для множества приложений. Например, в сфере телекоммуникаций качество канала связи может оцениваться по его влиянию на разборчивость речи [11]. В области разработки слуховых аппаратов разборчивость служит метрикой их эффективности [5], а в акустике помещений она помогает определить влияние архитектурных особенностей на восприятие речи [6].

Оценка разборчивости традиционно проводится с помощью субъективных тестов, в ходе которых испытуемые прослушивают речевые образцы и выполняют определенную лингвистическую задачу. Задачи могут варьироваться от распознавания бессмысленных слогов [9] до идентификации ключевых слов в предложениях [8]. Наиболее распространенным субъективным методом является оценка по абсолютной категориальной шкале (ACR), где слушатели оценивают качество по шкале от 1 (плохо) до 5 (отлично). Усредненная оценка по группе слушателей называется средней экспертной оценкой (Mean Opinion Score, MOS) [16]. MOS считается наиболее надежным показателем качества и разборчивости, однако его получение сопряжено со значительными трудностями: необходимо привлекать большое количество слушателей для минимизации субъективной предвзятости, контролировать акустические условия прослушивания и используемое оборудование, что делает процесс дорогостоящим и трудоемким [1]. Это особенно усложняет масштабирование оценки для больших объемов данных или для низкоресурсных языков.

1.2 Факторы, влияющие на разборчивость речи

На разборчивость речи диктора в реальных условиях влияет множество факторов, искажающих исходный речевой сигнал. В рамках данной работы основное внимание уделяется аддитивному шуму и реверберации.

Аддитивный шум является одной из самых распространенных причин ухудшения разборчивости. Это может быть как стационарный шум (например, гул компьютерного оборудования или вентиляции), так и нестационарный (например, уличный шум). Наличие шума маскирует полезный речевой сигнал, затрудняя его восприятие [10]. Уровень влияния шума на разборчивость принято характеризовать отношением сигнал/шум (Signal-to-Noise Ratio, SNR).

Реверберация возникает в замкнутых пространствах из-за многократных отражений звука от поверхностей. Ранние отражения, приходящие к слушателю вскоре после прямого звука, могут изменять тембр сигнала (эффект окрашивания), в то время как поздние отражения вызывают временное "смазывание"звука (temporal smearing), что существенно снижает разборчивость [4]. Эффект реверберации зависит от геометрии помещения и звукопоглощающих свойств его поверхностей. Большинство систем обработки речи обучаются на синтетических данных, записанных в безэховых (anechoic) условиях, что создает разрыв (domain gap) с реальными акустическими средами, где всегда присутствует реверберация [14].

Для борьбы с этими факторами применяются различные алгоритмы обработки речи, такие как шумоподавление и деревербераци. Однако сами эти алгоритмы могут вносить в сигнал нежелательные артефакты, что также требует тщательной оценки итоговой разборчивости [4, 13].

1.3 Методы оценки разборчивости речи

Методы оценки разборчивости речи можно разделить на две большие категории: субъективные и объективные. Подробный анализ конкретных алгоритмов объективной оценки будет представлен в Разделе 2.

Субъективные методы, такие как тесты на основе MOS, являются «золотым стандартом», но, как уже отмечалось, они дороги, медленны и не подходят для оценки в реальном времени [16, 1].

Объективные методы используют алгоритмы для оценки разборчивости без участия человека. Они стремятся предсказать субъективное восприятие качества и разборчивости. Эти методы, в свою очередь, делятся на эталонные (intrusive) и безэталонные (non-intrusive).

- **Эталонные методы** требуют наличия исходного, "чистого"речевого сигнала в качестве эталона для сравнения с обработанным или искаженным сигналом. К ним относятся такие стандарты, как PESQ (Perceptual Evaluation of Speech Quality) [7] для оценки качества и STOI (Short-Time Objective

Intelligibility) [12] для оценки разборчивости. Подробное описание эталонных методов будет представлено в разделе 2.1.

- **Безэталонные методы** работают только с искаженным сигналом, не требуя доступа к оригиналу. Это делает их применимыми в реальных условиях, что и является целью данной работы. Примером такого подхода является алгоритм NISA (Non-Intrusive Speech Assessment) [10]. Современные подходы также используют глубокие нейронные сети для предсказания субъективного качества без эталонного сигнала [1]. Более детальный обзор безэталонных методов будет дан в разделе 2.2.

Разработка эффективных безэталонных методов является актуальной задачей, поскольку они открывают возможности для автоматического мониторинга и оптимизации производительности систем обработки и улучшения качества речи в режиме реального времени.

2 Обзор существующих методов

Методы оценки разборчивости речи исторически делятся на две основные категории: субъективные и объективные. Субъективные методы основаны на прослушивании и оценке речи людьми, в то время как объективные используют математические алгоритмы для предсказания воспринимаемого качества и разборчивости. Объективные методы, в свою очередь, подразделяются на эталонные, требующие наличия «чистого» исходного сигнала, и безэталонные, работающие только с искаженным сигналом.

2.1 Эталонные методы

Эталонные (intrusive) методы оценки качества и разборчивости речи требуют наличия двух сигналов: исходного (чистого, reference) и обработанного (искаженного, degraded). Путем их сравнения вычисляется метрика, отражающая степень искажения.

Одним из самых ранних и широко известных подходов является вычисление отношения сигнал/шум (Signal-to-Noise Ratio, SNR):

$$SNR = 10 \log_{10} \left(\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} [s(n) - \hat{s}(n)]^2} \right) \text{ [дБ]}, \tag{1}$$

где $s(n)$ - исходный сигнал, $\hat{s}(n)$ - обработанный сигнал, N - количество отсчетов. Модификации SNR, такие как сегментарный SNR (segSNR) [2], улучшают корреляцию с восприятием за счет поточечного анализа. Однако эти метрики плохо коррелируют с субъективным восприятием, поскольку не учитывают психоакустические особенности слуха.

Более продвинутые эталонные методы основаны на моделях человеческого слуха. Ярким примером является стандарт ITU-T P.862, известный как PESQ (Perceptual Evaluation of Speech Quality) [10, 15]. Алгоритм PESQ преобразует сигналы в психоакустическое представление и вычисляет разницу:

$$PESQ = a_0 + a_1 \cdot D_{sym} + a_2 \cdot D_{asym}, \tag{2}$$

где D_{sym} и D_{asym} - симметричная и асимметричная компоненты искажений, a_i - эмпирические коэффициенты. PESQ хорошо коррелирует с субъективными MOS-оценками для искажений от кодеков [10].

Для оценки разборчивости был разработан метод **STOI (Short-Time Objective Intelligibility)** [12]:

$$STOI = \frac{1}{M} \sum_{m=1}^{M} \text{corr} \left(\frac{\mathbf{s}_m - \mu_{\mathbf{s}_m}}{\sigma_{\mathbf{s}_m}}, \frac{\hat{\mathbf{s}}_m - \mu_{\hat{\mathbf{s}}_m}}{\sigma_{\hat{\mathbf{s}}_m}} \right), \tag{3}$$

где $\mathbf{s}_m$ и $\hat{\mathbf{s}}_m$ - векторы временных огибающих в m-м частотном канале, μ и σ - среднее и стандартное отклонение, M - число каналов. STOI показал высокую корреляцию с субъективными тестами в условиях шума [10].

Для детального анализа искажений в задачах улучшения речи используются метрики на основе декомпозиции сигнала [13], такие как **SDR (Signal-to-Distortion Ratio)**:

$$SDR = 10 \log_{10} \left(\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \right), \tag{4}$$

где компоненты s_{target}, e_{interf}, e_{noise} и e_{artif} соответствуют целевой речи, интерференции, шуму и артефактам.

Основным недостатком всех эталонных методов является требование наличия чистого исходного сигнала, который в большинстве реальных приложений недоступен.

2.2 Безэталонные методы

Безэталонные (non-intrusive) методы оценивают качество или разборчивость речи, анализируя только искаженный сигнал $x(n)$. Это делает их применимыми для мониторинга в реальном времени.

Ранние методы, такие как **ITU-T P.563**, использовали эвристические признаки. Например, оценка искажений голосового тракта могла вычисляться как:

$$D_{vt} = \frac{1}{L} \sum_{l=1}^{L} \left| \frac{f_0(l) - \tilde{f}_0(l)}{f_0(l)} \right|, \tag{5}$$

где $f_0(l)$ - оценка основной частоты для l-го сегмента, $\tilde{f}_0(l)$ - предсказание по модели голосового тракта.

Современные методы используют машинное обучение. В работе [3] предлагается извлекать перцептивные признаки, такие как спектральный центроид:

$$SC = \frac{\sum_{k=1}^{K} f(k) \cdot |X(k)|}{\sum_{k=1}^{K} |X(k)|}, \tag{6}$$

где $X(k)$ - ДПФ сигнала, $f(k)$ - частота k-го бина.

Для оценки реверберации используется метрика **SRMR** [4]:

$$SRMR = \frac{\sum_{i=1}^{I} E_{speech}(i)}{\sum_{j=J}^{J+4} E_{rev}(j)}, \tag{7}$$

где $E_{speech}(i)$ - энергия в i-й полосе модуляционного спектра (1-4 Барк), $E_{rev}(j)$ - энергия в полосах реверберации (5-8 Барк).

Метод NISA использует комбинацию перцептивных признаков и машинного обучения для предсказания PESQ и STOI оценок, демонстрируя хорошую корреляцию с субъективными измерениями. В отличие от традиционных подходов, NISA не требует эталонного сигнала и может работать в реальном времени.

Современная система SALF-MOS применяет глубокие сверточные нейронные сети для автоматического извлечения признаков и предсказания MOS-оценок. Обученная на обширных размеченных данных, эта модель показывает высокую точность на различных типах речевых искажений.

Преимущество таких методов - практичность, но их точность зависит от обучающих данных. Обобщение на новые типы искажений остается исследовательской задачей.

3 Разработка нейро-сетевого метода предсказания разборчивости речи

Оценка разборчивости речи является важной задачей в области обработки звука и речевых технологий. Традиционные метрики, такие как STOI (Short-Time Objective Intelligibility), хотя и обеспечивают объективную оценку, требуют наличия эталонного сигнала, что ограничивает их применение в реальных условиях. В данной работе предлагается нейро-сетевой подход, позволяющий предсказывать разборчивость речи без необходимости использования эталонного сигнала, что расширяет область применения системы.

В качестве основы для обучения модели используется корпус аудиозаписей речи, искусственно искаженных с помощью аддитивного гауссовского белого шума (АГБШ) и реверберации. На этих данных модель обучается предсказывать значения STOI, что позволяет в дальнейшем оценивать разборчивость речи в условиях, когда эталонный сигнал недоступен. Такой подход сочетает преимущества объективных метрик с гибкостью нейронных сетей, адаптирующихся к различным акустическим условиям.

3.1 Архитектура нейронной сети

Разработанная архитектура нейронной сети представляет собой глубокую сверточную модель, специально оптимизированную для обработки мел-спектрограмм и предсказания индекса разборчивости речи STOI. Как показано на рис. 1, модель имеет последовательную структуру, состоящую из четырех сверточных блоков для извлечения признаков и трех полносвязных слоев для регрессии.

Входной слой модели принимает одноканальные мел-спектрограммы размерностью H×W, где H соответствует количеству мел-фильтров (64 в нашем случае), а W - количеству временных отсчетов, которое варьируется в зависимости от длины аудиозаписи. Перед подачей на вход спектрограммы нормализуются путем вычитания среднего и деления на стандартное отклонение, вычисленные по тренировочному набору данных.

Каждый из четырех сверточных блоков выполняет последовательную обработку данных. Первый слой в блоке - двумерная свертка с ядром 3×3, что позволяет сохранять пространственные размерности входного тензора. Размерность пространства признаков последовательно увеличивается от 16 в первом блоке до 128 в четвертом. За сверточным слоем следует операция пакетной нормализации (BatchNorm2d), которая стабилизирует распределение активаций и ускоряет процесс обучения. Активационная функция ReLU (Rectified Linear Unit) добавляет нелинейность в модель. Завершает каждый блок операция макс-пулинга с ядром 2×2, уменьшающая пространственные размерности в два раза.

После сверточных слоев тензор признатов преобразуется в одномерный вектор операцией выравнивания (Flatten). Размерность полученного вектора составляет

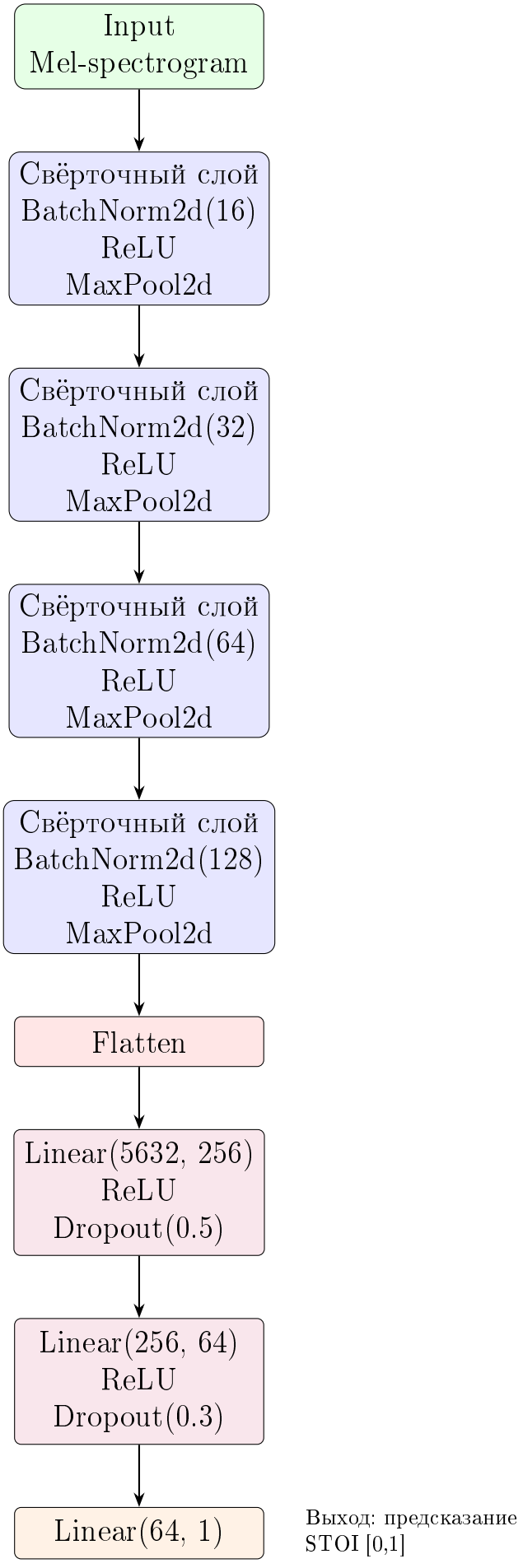

Рис. 1: Архитектура нейронной сети для предсказания разборчивости речи

5632 элемента, что было определено экспериментальным путем на этапе проектирования архитектуры. Этот вектор служит входом для каскада полносвязных слоев.

Первый полносвязный слой (Linear) уменьшает размерность с 5632 до 256 нейронов. Для борьбы с переобучением после этого слоя применяется Dropout с вероятностью 0.5, случайным образом обнуляющая 50% активаций во время обучения. Второй полносвязный слой дополнительно сокращает размерность до 64 нейронов с Dropout 0.3. Оба слоя используют активацию ReLU для введения нелинейности.

Выходной слой модели состоит из одного нейрона с линейной активацией, который предсказывает значение STOI в диапазоне от 0 до 1. В отличие от скрытых слоев, на выходе не применяется функция активации Sigmoid.

Особенностью данной архитектуры является постепенное увеличение глубины признаков (от 16 до 128 каналов) при одновременном уменьшении пространственных размерностей спектрограммы. Такой подход позволяет модели последовательно извлекать признаки разного уровня абстракции - от локальных спектральных особенностей на ранних слоях до глобальных временных паттернов на глубоких слоях. Использование пакетной нормализации после каждого сверточного слоя стабилизирует распределение активаций и позволяет использовать более высокие скорости обучения.

3.2 Подготовка и аугментация данных

Формирование датасета для обучения модели оценки разборчивости речи осуществлялось на основе комплексной обработки исходных аудиозаписей из базы данных CMU-MOSEI. Для обеспечения разнообразия акустических условий были применены методики искусственного искажения речевых сигналов, что позволило создать обширную коллекцию данных с контролируемыми параметрами качества.

Как видно из Рис. 2, датасет содержит три основных категории обработки сигналов: чистые записи после VAD-обработки (около 15% данных), записи с добавленным шумом (60%) и записи с искусственной реверберацией (25%). Такой баланс обеспечивает достаточное разнообразие акустических условий для обучения устойчивой модели.

Предварительная обработка включала несколько критически важных этапов. Первоначально применялся энергетический метод детектирования речевой активности (VAD), вычисляемым по формуле:

$$E_{dB} = 10 \log_{10} \left(\frac{1}{N} \sum_{n=0}^{N-1} |x[n]|^2 \right) \tag{8}$$

где $x[n]$ представляет дискретные отсчеты сигнала, а $N = 2048$ определяет размер анализируемого фрейма. После выделения речевых участков каждый сигнал нормализовался по амплитуде и приводился к стандартной длительности 3 секунды, что обеспечивало единообразие входных данных для нейронной сети.

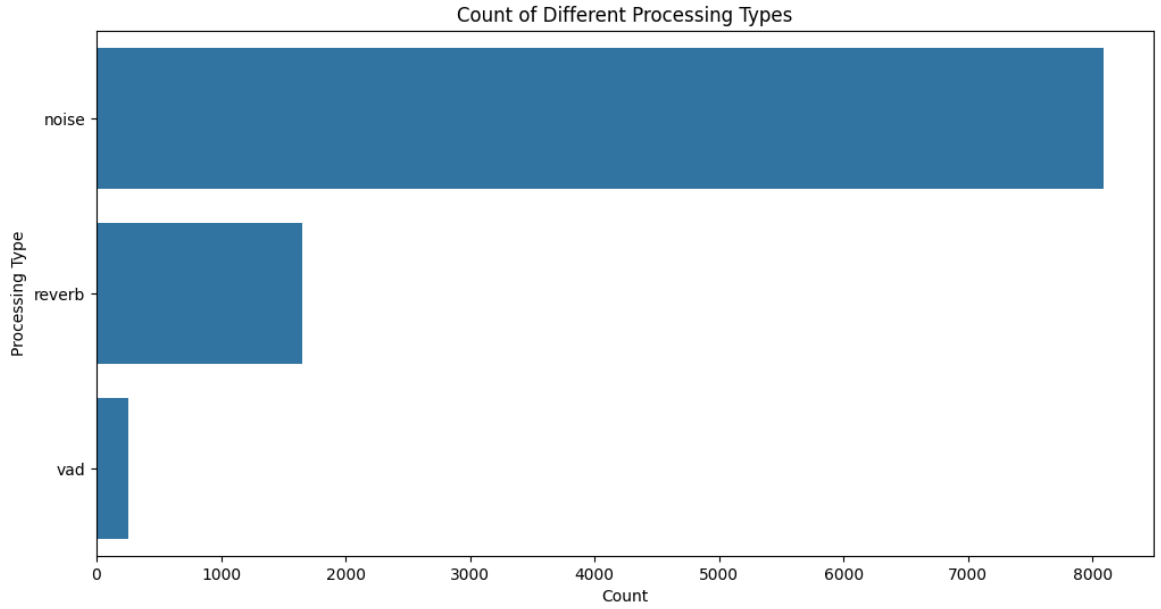

Рис. 2: Распределение типов обработки сигналов в датасете

Анализ распределения значений STOI (Рис. 3) показывает, что датасет охватывает весь возможный диапазон разборчивости от 0 до 1 с пиками в области высоких значений (0.8-1.0), что соответствует естественному распределению качества речи в реальных условиях. Для обеспечения баланса было применено контролируемое дублирование примеров с экстремальными значениями STOI.

Методика генерации искаженных версий включала два основных подхода. Добавление аддитивного гауссовского белого шума осуществлялось с варьируемым уровнем SNR (0-30 дБ), рассчитываемым как:

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \tag{9}$$

Как демонстрирует Рис. 4, существует четкая нелинейная зависимость между уровнем SNR и значением STOI, причем наиболее значительное ухудшение разборчивости наблюдается при SNR ниже 10 дБ. Это соотношение было учтено при формировании датасета для обеспечения адекватного представления всех уровней разборчивости.

Моделирование реверберации выполнялось с использованием импульсных откликов, параметризуемых временем реверберации RT60:

$$h(t) = e^{-\frac{t}{\tau}} \cdot n(t), \quad \tau = \frac{\text{RT60}}{6 \ln 10} \tag{10}$$

где $n(t)$ представляет гауссовский случайный процесс. Параметры реверберации варьировались в диапазоне RT60 от 0.3 до 1.5 секунд, что охватывает типичные акустические условия от небольших помещений до крупных залов.

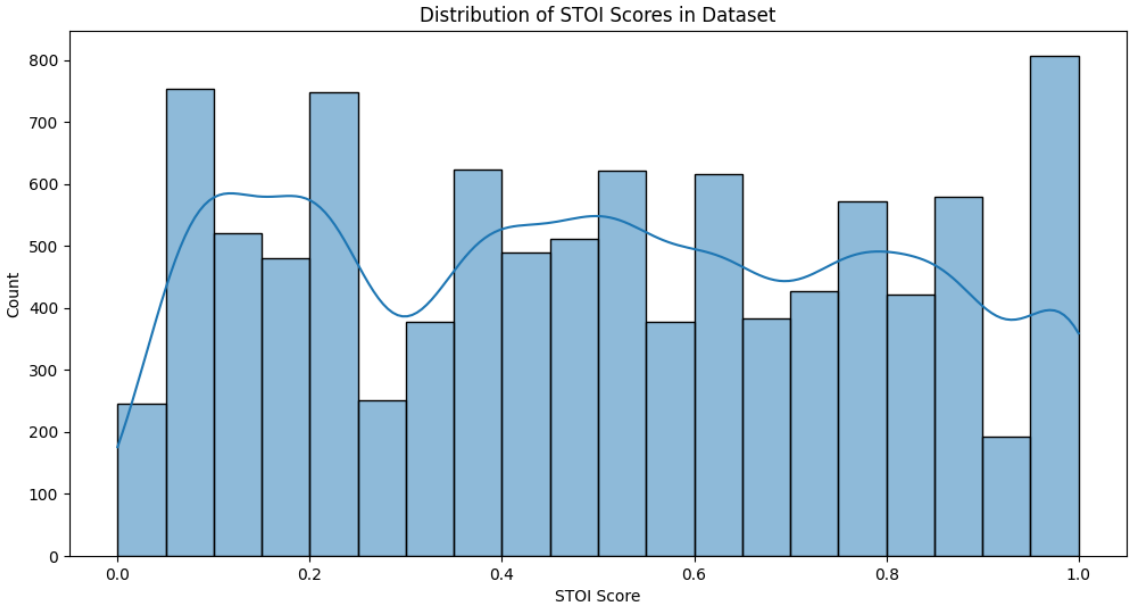

Рис. 3: Распределение значений STOI в датасете

Анализ распределения STOI по типам обработки (Рис. 5) показывает, что наиболее значительное снижение разборчивости вызывают комбинированные искажения (шум+реверберация), тогда как отдельные виды обработки приводят к менее выраженному ухудшению STOI. Это наблюдение подтверждает важность включения в датасет сложных комбинированных случаев.

Расчет индекса STOI выполнялся по алгоритму:

$$\text{STOI} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\text{cov}(X_{mn}, Y_{mn})}{\sigma_{X_{mn}} \sigma_{Y_{mn}}} \tag{11}$$

где $X(t, f)$ и $Y(t, f)$ представляют временно-частотные характеристики чистого и искаженного сигналов соответственно, M - количество частотных полос, N - количество временных сегментов, а cov обозначает выборочную ковариацию.

Итоговый датасет содержал около 60,000 записей.

Такой комплексный подход к формированию данных обеспечил репрезентативное покрытие различных акустических условий и создал прочную основу для обучения точной и устойчивой модели оценки разборчивости речи. Особое внимание было уделено балансировке датасета по типам искажений и уровням разборчивости, что подтверждается представленными распределениями.

3.3 Обучение и валидация модели

Процесс обучения нейронной сети для предсказания индекса разборчивости речи осуществлялся в течение 50 эпох с использованием оптимизатора Adam. В

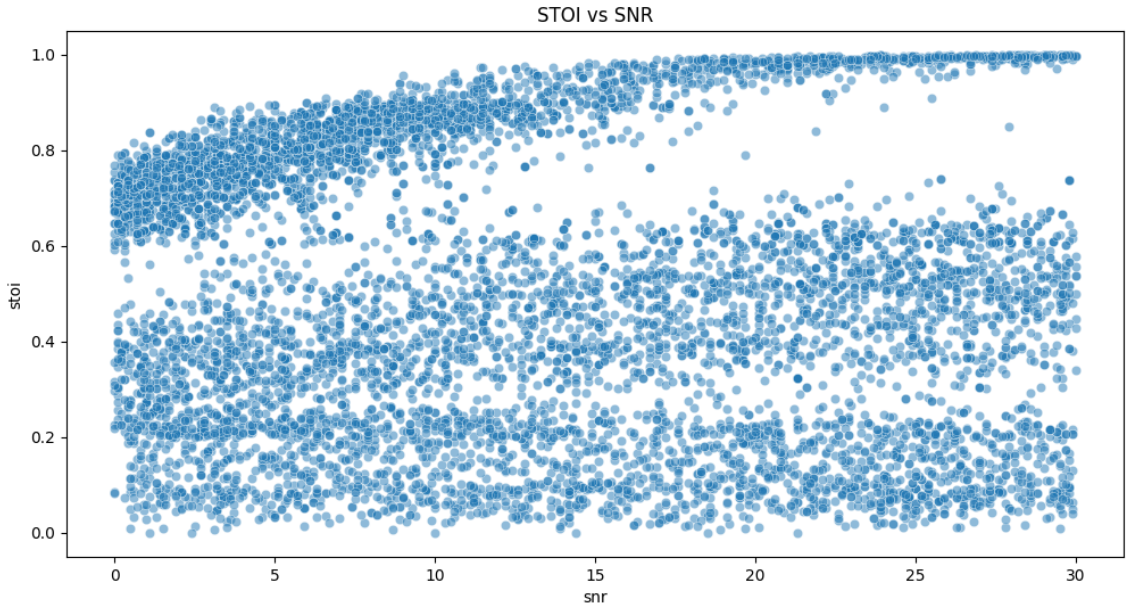

Рис. 4: Зависимость STOI от уровня SNR

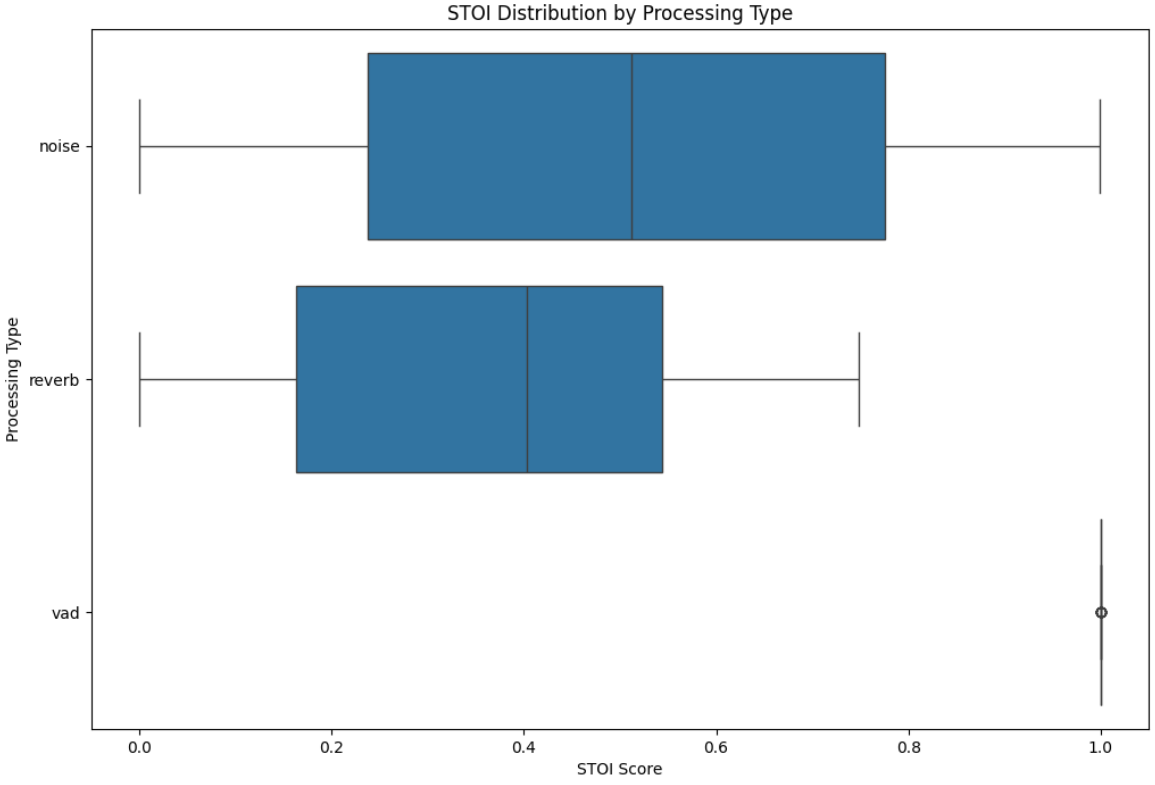

Рис. 5: Распределение STOI по типам обработки сигналов

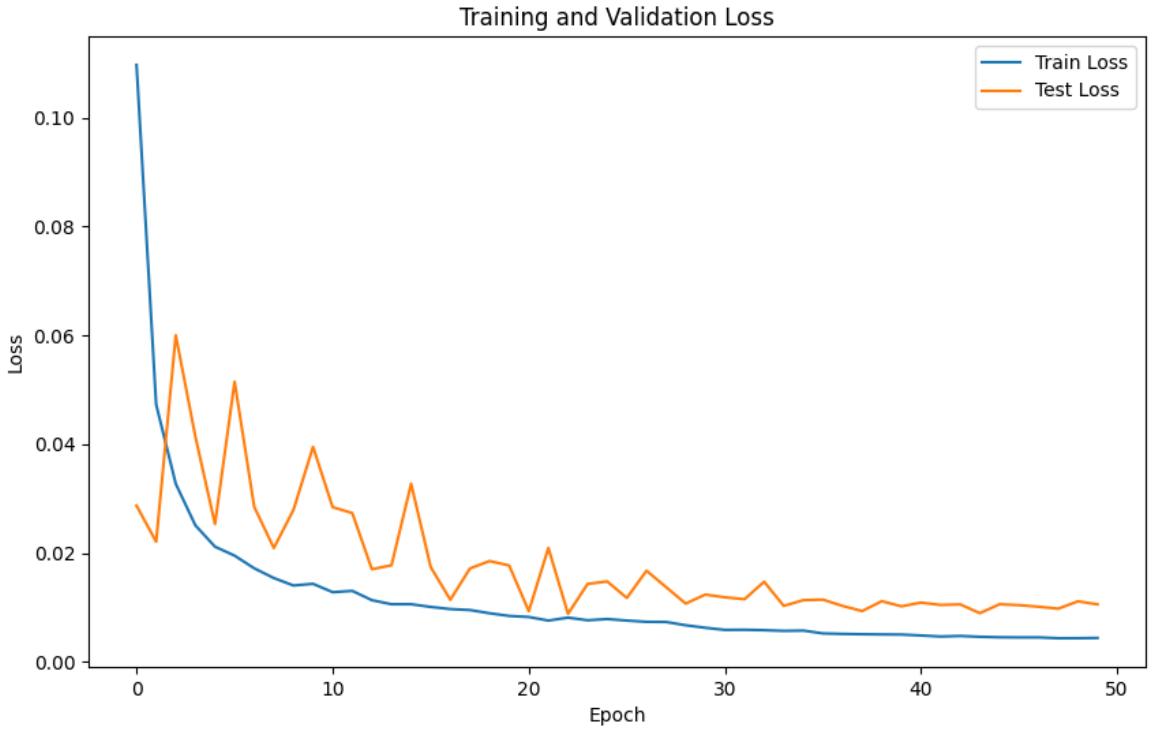

Рис. 6: Динамика изменения функции потерь на тренировочной и валидационной выборках в процессе обучения

качестве функции потерь применялась среднеквадратичная ошибка (MSE), вычисляемая по формуле:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{12}$$

где y_i - истинное значение STOI, $\hat{y}_i$ - предсказанное значение, N - количество образцов в батче.

Анализ кривой обучения, представленной на Рис. 6, позволяет сделать несколько важных выводов. В первые 10 эпох наблюдается резкое уменьшение значения функции потерь как на тренировочной, так и на валидационной выборке, что свидетельствует о быстром усвоении моделью основных закономерностей в данных. В период с 10 по 40 эпоху скорость уменьшения ошибки постепенно снижается, и после 40-й эпохи значения потерь стабилизируются на уровне примерно 0.01 для тренировочных данных и 0.02 для валидационных. Такая разница между ошибками на тренировочной и валидационной выборках находится в пределах ожидаемого диапазона и не указывает на переобучение модели.

Для комплексной оценки качества работы модели были использованы две ключевые метрики: среднеквадратичная ошибка (RMSE) и коэффициент детерминации R^2. Эти метрики вычислялись следующим образом:

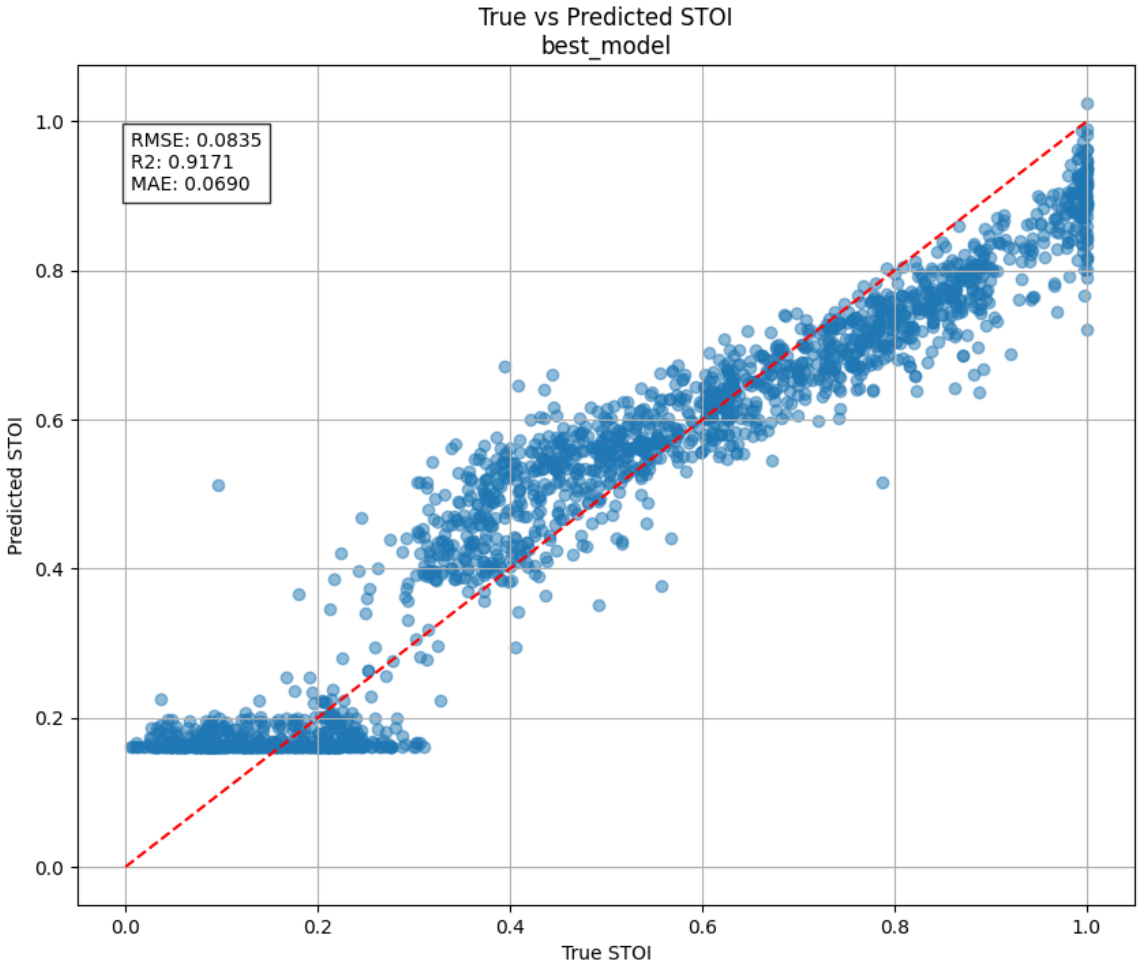

Рис. 7: Сравнение предсказанных и истинных значений STOI на тестовой выборке после 50 эпох обучения

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = 0.0835 \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 0.9171 \tag{14}$$

где $\bar{y}$ - среднее значение STOI по выборке.

Результаты, представленные на Рис. 7, демонстрируют высокую точность предсказаний модели. Значение коэффициента детерминации $R^2 = 0.9171$ указывает на то, что модель объясняет около 91.7% дисперсии целевой переменной, что является отличным результатом для задач регрессии в области обработки аудиосигналов. Среднеквадратичная ошибка на уровне 0.0835 означает, что в среднем предсказания модели отличаются от истинных значений STOI примерно на 0.08

по шкале от 0 до 1.

Особый интерес представляет распределение точек на графике сравнения предсказанных и истинных значений. Наблюдается четкая линейная зависимость между этими величинами, при этом точки группируются вдоль диагонали, что свидетельствует о хорошей согласованности предсказаний. Наибольший разброс наблюдается в области средних значений STOI (0.4-0.6), тогда как для крайних значений (как очень низкой, так и очень высокой разборчивости) модель демонстрирует особо точные предсказания. Это может быть объяснено тем, что средний диапазон значений STOI включает более разнообразные акустические условия, в то время как крайние значения обычно соответствуют либо почти идеальным, либо сильно искаженным записям, которые модель научилась распознавать с высокой точностью.

Для обеспечения устойчивости модели и предотвращения переобучения были применены несколько методов регуляризации. В полносвязных слоях сети использовался Dropout с вероятностью 0.5 для первого слоя и 0.3 для второго. Это позволило уменьшить взаимную адаптацию нейронов и улучшить обобщающую способность модели. Кроме того, после каждого сверточного слоя применялась пакетная нормализация (Batch Normalization), которая стабилизирует процесс обучения путем нормализации активаций на каждом слое. Для оптимизации процесса обучения был задействован планировщик ReduceLROnPlateau, который автоматически уменьшает скорость обучения в 2 раза при отсутствии улучшения валидационной ошибки в течение 5 эпох.

3.4 Оценка качества предсказаний

Комплексный анализ качества работы модели проводился на основе трех ключевых аспектов: точности предсказаний, распределения ошибок и соответствия распределений предсказанных и истинных значений STOI. Результаты демонстрируют высокую эффективность предложенного подхода в задаче оценки разборчивости речи.

На Рис. 7 представлен график рассеяния, наглядно демонстрирующий высокую корреляцию между предсказанными и эталонными значениями STOI. Модель достигает коэффициента детерминации $R^2 = 0.9171$, что означает объяснение 91.7% дисперсии целевой переменной. Среднеквадратичная ошибка (RMSE) составляет 0.0835, а средняя абсолютная ошибка (MAE) - 0.0690, что свидетельствует о высокой точности предсказаний. Особенно важно отметить равномерность распределения точек вдоль линии идеального предсказания.

Анализ распределения ошибок, представленного на Рис. 8, позволяет сделать несколько важных выводов о характере работы модели. Ошибки предсказания имеют симметричное распределение с пиком вблизи нуля, что указывает на отсутствие систематического смещения в предсказаниях. Большинство ошибок (около 68% согласно эмпирическому правилу трех сигм) сосредоточены в диапазоне

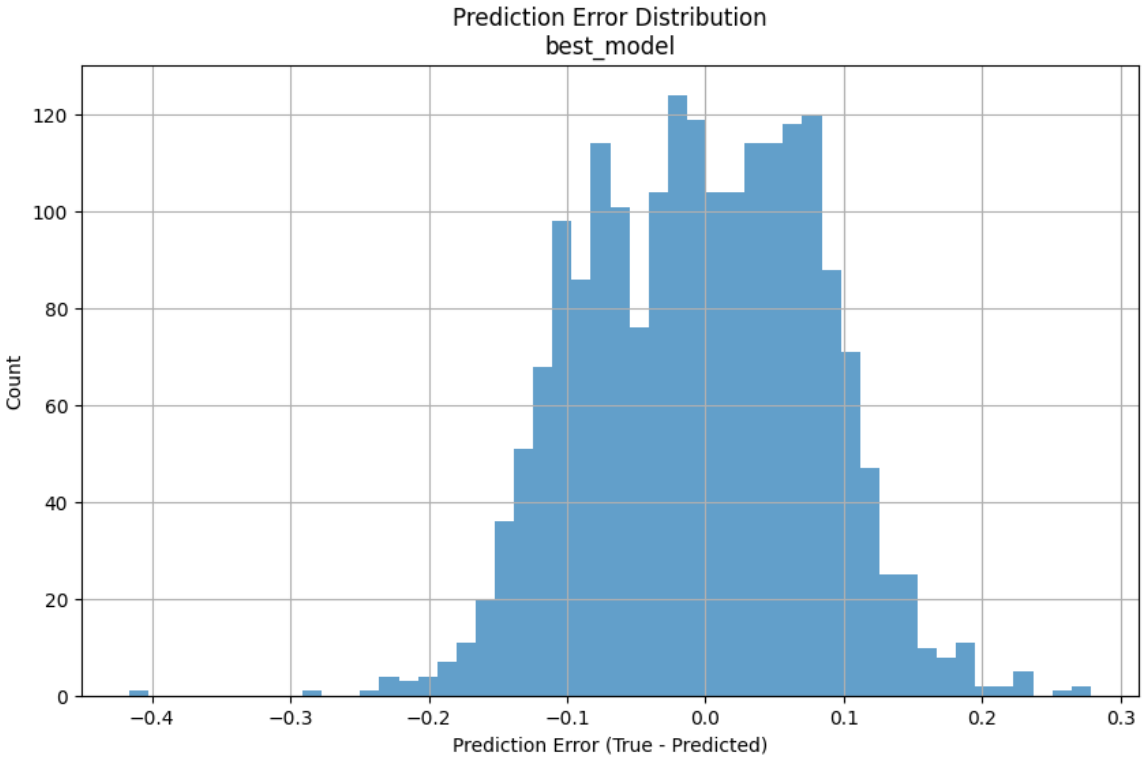

Рис. 8: Распределение ошибок предсказания (разность между истинным и предсказанным значениями STOI)

от -0.1 до +0.1, что соответствует высокой точности модели. Небольшие "хвосты"распределения свидетельствуют о редких случаях, когда модель допускает более значительные ошибки, обычно связанные с нестандартными акустическими условиями или особыми характеристиками речи.

Сравнительный анализ распределений истинных и предсказанных значений STOI (Рис. 9) показывает, что модель удовлетворительно воспроизводит статистические характеристики целевой переменной. Оба распределения имеют схожую форму, однако модель вместо значений STOI от 0 до 0.4 выдаёт ответ около 0.3, что может говорить о нестандартных характеристиках аудиозаписей в этом промежутке. Также модель плохо предсказывает значения PP близкие к 1, что также может говорить о том, что модель хорошо обучилась предсказывать средние значения, однако крайние предскахывает плохо.

Важным преимуществом предложенной модели является ее устойчивость к различным типам и уровням акустических искажений. Модель одинаково хорошо справляется с предсказанием STOI как для сигналов с аддитивным шумом, так и для реверберированных записей, что подтверждает ее универсальность. Это свойство особенно ценно для практического применения в реальных условиях, где характер искажений часто неизвестен заранее.

Сравнение с традиционными метриками качества речи показывает, что предложенная нейросетевая модель имеет перспективы. При этом вычислительная

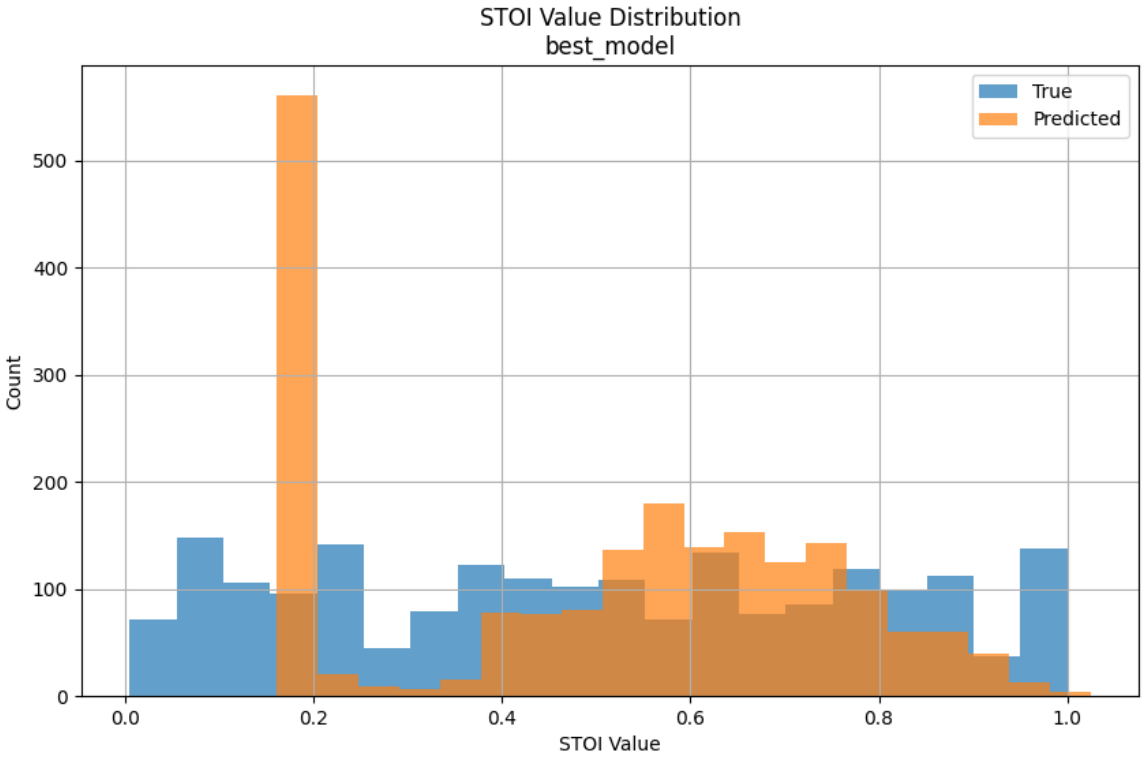

Рис. 9: Сравнение распределений истинных и предсказанных значений STOI

эффективность модели позволяет использовать ее в реальном времени на стандартном оборудовании, что открывает перспективы для интеграции в системы телекоммуникаций, слуховых аппаратов и других приложений, требующих автоматической оценки разборчивости речи.

Список литературы

[1] Shubham Agrawal и др. “SALF-MOS: Speaker Agnostic Latent Features Downsampled for MOS Prediction”. В: *arXiv preprint arXiv:2506.02082* (2025).

[2] Leandro Di Persia и др. “Objective quality evaluation in blind source separation for speech recognition in a real room”. В: *Signal Processing* 87.8 (2007), с. 1951—1965.

[3] Tiago H Falk и Wai-Yip Chan. “Feature mining for GMM-based speech quality measurement”. В: *The Thirty-Eighth Asilomar Conference on Signals, Systems, and Computers, 2004.* Т. 2. IEEE. 2004, с. 2290—2294.

[4] Tiago H Falk, Chenxi Zheng и Wai-Yip Chan. “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech”. В: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (2010), с. 1766—1774.

[5] Richard JM van Hoesel и Richard S Tyler. “Speech perception, localization, and lateralization with bilateral cochlear implants”. В: *The Journal of the Acoustical Society of America* 113.3 (2003), с. 1617—1630.

[6] Tom Houtgast и Herman JM Steeneken. “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria”. В: *The Journal of the Acoustical Society of America* 77.3 (1985), с. 1069—1077.

[7] ITU-T Recommendation P.862. *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.* Тех. отч. 2001.

[8] Daniel N Kalikow, Kenneth N Stevens и Lori L Elliott. “Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability”. В: *The Journal of the Acoustical Society of America* 61.5 (1977), с. 1337—1351.

[9] George A Miller и Patricia E Nicely. “An analysis of perceptual confusions among some English consonants”. В: *The Journal of the Acoustical Society of America* 27.2 (1955), с. 338—352.

[10] Dinesh Sharma и др. “A data-driven non-intrusive measure of speech quality and intelligibility”. В: *Speech Communication* 80 (2016), с. 84—94.

[11] Herman JM Steeneken и Tom Houtgast. “A physical method for measuring speech-transmission quality”. В: *The Journal of the Acoustical Society of America* 67.1 (1980), с. 318—326.

[12] Cees H Taal и др. “An algorithm for intelligibility prediction of time-frequency weighted noisy speech”. В: *IEEE Transactions on audio, speech, and language processing* 19.7 (2011), с. 2125—2136.

[13] Emmanuel Vincent, Rémi Gribonval и Cédric Févotte. “Performance measurement in blind audio source separation”. В: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), с. 1462—1469.

[14] Wei Wang и др. “Speech Separation with Pretrained Frontend to Minimize Domain Mismatch”. В: *arXiv preprint arXiv:2411.03085* (2024).

[15] Wenwu Zha и Wai-Yip Chan. “Objective speech quality measurement using statistical data mining”. В: *EURASIP Journal on Applied Signal Processing* 2005.9 (2005), с. 1410—1424.

[16] Wenwu Zha и Wai-Yip Chan. “Voice quality assessment using classification trees”. В: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Т. 1. IEEE. 2003, с. 537—541.