Сравнение нейросетевых архитектур для распознавания русской речи с иностранным акцентом

А.А. Волкова, Е.В. Дружинская

***Аннотация*—Системы автоматического распознавания речи активно развиваются благодаря широкому применению методов глубокого обучения. Однако при работе с русской речью, произносимой носителями других языков, большинство алгоритмов сталкиваются со снижением точности. Акцентная специфика влияет на длительность звуков, артикуляционные переходы и просодические характеристики, что затрудняет корректное выделение акустических признаков и последующую транскрибацию.**

В научной литературе описано множество подходов к обработке речи, включая глубокие, сверточные и рекуррентные модели, а также гибридные архитектуры, использующие механизмы внимания. Каждая из них по-разному реагирует на вариативность произношения и степень выраженности акцента.

Учитывая эти особенности, работа направлена на изучение поведения различных нейросетевых архитектур при распознавании русской речи с иностранным акцентом и на анализ их результатов на корпусе собственных записей. Оценка проводится по метрикам WER, CER, Accuracy, что позволяет выявить модели, демонстрирующие наибольшую устойчивость к акцентным искажениям и способные работать с ограниченными и неоднородными данными.

***Ключевые слова*—распознавание речи, акцентированная речь, нейросетевые архитектуры, обработка речи, глубокое обучение.**

I. Введение

Автоматическое распознавание речи постепенно становится важной частью повседневных технологий: голосовых интерфейсов, образовательных платформ и систем человек-машина. Качество таких систем зависит от того, насколько алгоритмы способны учитывать особенности живой речи, включая темп, интонацию и вариативность артикуляции. Особенно заметные трудности возникают при работе с русской речью, произносимой носителями других языков. Иностранный акцент приводит к систематическим изменениям звучания: меняется длительность фонем, нарушаются привычные артикуляционные переходы и искажается просодический рисунок. Эти особенности отражаются на спектральной структуре сигнала и усложняют выделение акустических признаков, что нередко приводит к снижению точности транскрипции даже в современных системах, обученных на больших наборах данных.

Русский язык обладает сложной фонетической системой, поэтому даже слабый акцент способен привести к значимым отклонениям от стандартных моделей звучания. Системы, обученные преимущественно на речи носителей, как правило, демонстрируют более высокие уровни ошибок при работе с иностранцами, изучающими русский язык. Подобные результаты выявляют ограничения существующих решений и подчёркивают необходимость анализа их устойчивости к акцентным искажениям.

Ранее автором был выполнен анализ особенностей акцентированной русской речи у иностранных студентов и рассмотрены модели, потенциально применимые для её автоматического распознавания [1]. Настоящее исследование развивает данное направление и сосредоточено на экспериментальной оценке различных нейросетевых архитектур в условиях ограниченного корпуса акцентированной речи.

Современные системы распознавания речи используют широкий спектр архитектур — от сверточных, рекуррентных и глубоких моделей до гибридных решений с механизмами внимания. Они различаются по подходам к обработке временной структуры сигнала, способам извлечения признаков и устойчивости к вариативности произношения. Однако степень их чувствительности к акцентным отклонениям всё ещё изучена недостаточно, особенно применительно к русской речи.

Акцентированная речь отличается высокой неоднородностью: записи варьируются по темпу, интонации, качеству дикции, условиям записи и уровню владения языком. Эти факторы усложняют задачу распознавания, особенно при ограниченном объёме данных. В связи с этим представляет интерес исследование поведения различных архитектур в таких условиях и оценка их устойчивости к фонетическим вариациям.

II. Архитектуры нейронных сетей для распознавания речи

Развитие методов глубокого обучения привело к появлению большого числа архитектур, используемых для анализа акустического сигнала и автоматического распознавания речи. Эти модели отличаются принципами обработки временных и спектральных

Статья получена 1 декабря 2025.

А.А. Волкова, Уфимский государственный нефтяной технический университет, Уфа, Россия (e-mail: nastya.volkova.me@gmail.com).

Е.В. Дружинская, Уфимский государственный нефтяной технический университет, Уфа, Россия (e-mail: d.alena1806@gmail.com).

характеристик речи, что определяет их устойчивость к вариативности произношения, в том числе к акцентным искажениям. В данном разделе рассматриваются архитектуры, участвующие в сравнительном анализе: глубокие нейронные сети (DNN), сверточные сети (CNN), рекуррентные сети (RNN), модификации RNN с механизмами долговременной памяти (LSTM),stacked RNN, Conformer и свёрточно-рекуррентная нейронная сеть (CRNN).

A. Глубокие нейронные сети (DNN)

Глубокие нейронные сети представляют собой разновидность искусственных нейронных сетей, содержащих несколько скрытых слоев между входным и выходным слоями (рис. 1). Благодаря многослойной структуре DNN способны моделировать сложные нелинейные зависимости и эффективно обрабатывать большие объемы данных. Входной слой получает необработанные данные, где каждый нейрон представляет отдельный признак входного набора данных. Скрытые слои состоят из нейронов, связанных с нейронами соседних слоев, и выполняют нелинейные преобразования входных данных, применяя функции активации. Выходной слой предоставляет окончательные предсказания модели [2].

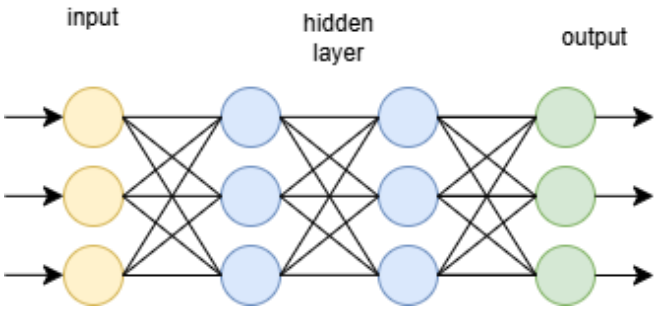

Рисунок 1 - архитектура DNN

DNN хорошо моделируют статические зависимости между признаками, однако не учитывают временную структуру акустического сигнала. При работе с акцентированной речью это приводит к трудностям в обработке изменяющейся длительности звуков и нарушенной просодики.

B. Сверточные нейронные сети (CNN)

Сверточные нейронные сети находят применение в обработке двумерных данных и широко используются в области компьютерного зрения. Их архитектура включает несколько сверточных слоев, соединенных с полносвязными нейронами (рис. 2).

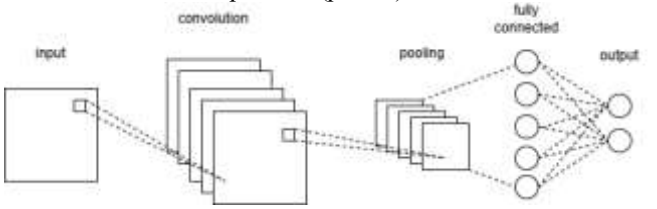

Рисунок 2 - архитектура CNN

По сравнению с другими глубокими моделями, CNN демонстрируют высокую эффективность при обработке изображений и акустических данных [3].

CNN выделяют устойчивые спектральные признаки речи, сглаживая шумы и локальные вариации. Это делает их полезными для обработки акцентированной речи, где спектральные свойства сигнала могут смещаться. Однако CNN не моделируют длинную временную зависимость.

C. Рекуррентные нейронные сети (RNN)

Рекуррентные нейронные сети представляют собой класс нейронных сетей, специально разработанных для обработки последовательных данных. В отличие от традиционных нейронных сетей, которые обрабатывают данные независимо, RNN учитывают временные зависимости между элементами последовательности, что делает их особенно эффективными для задач, связанных с временными рядами, текстами, речью и другими последовательностями [3]. Рекуррентные нейронные сети характеризуются наличием циклов в своей структуре, что позволяет им сохранять информацию о предыдущих состояниях и использовать её для обработки текущих данных (рис. 3). Это свойство делает RNN мощным инструментом для анализа последовательностей, где порядок элементов имеет значение [4].

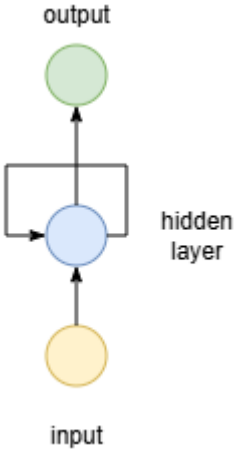

Рисунок 3 - архитектура RNN

RNN способны улавливать закономерности в динамике речи, что делает их более устойчивыми к акцентным особенностям, связанным с изменением длительности и скорости произнесения.

D. Сети с долговременной памятью (LSTM)

LSTM представляют собой модификацию рекуррентных сетей, разработанную для устойчивой обработки длинных последовательностей. Основным элементом архитектуры является **ячейка памяти**, которая содержит внутреннее состояние и набор обучаемых вентилей [5]. Такая структура позволяет контролировать, какие данные сохраняются, а какие — отбрасываются, предотвращая проблему исчезающих градиентов (рис. 4).

Работа LSTM основывается на трех управляющих механизмах:

- *механизм забывания* определяет, какая часть предыдущего состояния будет сохранена;
- *механизм обновления* контролирует, какие новые значения могут быть внесены в память;

- *механизм формирования выхода* задает, какая информация будет передана на следующий временной шаг [6].

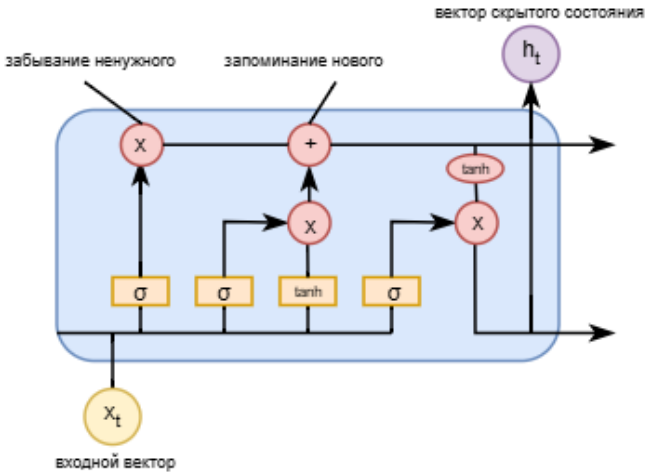

Рисунок 4 - архитектура LSTM

За счет управления потоком информации LSTM лучше сохраняют длительные зависимости и справляются с вариативностью акцентированной речи, где длительность фонем и переходы между ними могут существенно отличаться от норм речи носителей.

E. Stacked RNN

Stacked RNN представляют собой расширение классической архитектуры рекуррентных сетей, в которой несколько RNN-слоёв размещаются последовательно друг над другом. В отличие от простой однослойной RNN, использующей только одно скрытое состояние, стековая архитектура формирует многоуровневую систему обработки последовательности (рис. 5). Выход каждого RNN-слоя передаётся на вход следующему слою, благодаря чему модель способна извлекать более сложные абстракции [7].

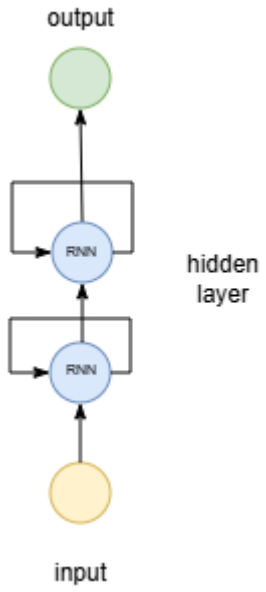

Рисунок 5 - архитектура stacked RNN

Стековая RNN особенно эффективна для обработки данных, в которых присутствуют вложенные временные структуры и многослойная вариативность. В задаче распознавания русской речи с иностранным акцентом такая архитектура помогает учитывать как быстрые изменения в произношении отдельных фонем, так и длительные контекстные зависимости, возникающие из-за нерегулярного темпа, изменённых артикуляционных переходов и специфики интонации.

F. Архитектура Conformer

Архитектура Conformer была разработана как сочетание механизмов самовнимания и сверточных преобразований, что позволяет эффективно учитывать как глобальные, так и локальные особенности речевого сигнала (рис. 6). В отличие от традиционных Transformer-моделей, Conformer дополнительно включает сверточный модуль, который позволяет учитывать локальные зависимости, характерные для акустического сигнала. Сверточные слои эффективно выделяют локальные особенности звука, воспринимая его структурные нюансы, а блоки самовнимания моделируют глобальные зависимости, что обеспечивает глубокое и всестороннее понимание речевых данных [8].

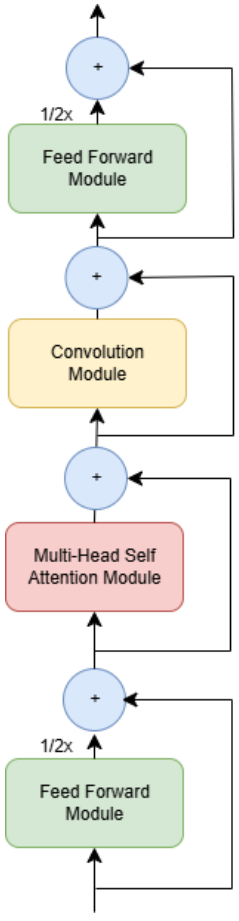

Рисунок 6 - архитектура Transformer

Архитектура Conformer состоит из следующих компонентов:

- *Feed-Forward блоков*, обеспечивающих нелинейные преобразования;
- *Multi-Head Self-Attention (MHSA)*, моделирующего глобальные зависимости между фрагментами речи;
- *Convolution Module*, обрабатывающего локальные временные структуры [9].

Такая комбинация делает Conformer особенно эффективным при работе с акцентированной речью. Сверточная часть сглаживает локальные искажения, а механизм внимания учитывает общую структуру произнесённой фразы.

G. Свёрточно-рекуррентная нейронная сеть (CRNN)

CRNN применяются для обработки последовательных данных, где важно учитывать как локальные особенности сигнала, так и его временную динамику. В контексте распознавания речи такая архитектура удобна тем, что соединяет преимущества свёрточных слоёв, выделяющих устойчивые акустические признаки, и рекуррентных слоёв, отвечающих за анализ временной структуры звучания (рис. 7).

Работа CRNN определяется двумя ключевыми компонентами:

- *Свёрточная сеть (CNN)* принимает на вход mel-спектрограммы, содержащие информацию о частотном составе речи во времени. Свёрточные слои выделяют локальные акустические паттерны: переходы между фонемами, формантные структуры, шумовые артефакты, особенности артикуляции, возникающие при иностранном акценте. Пулинг выполняется по частотной оси, что уменьшает размерность признаков, сохраняя временную последовательность. Это позволяет модели извлекать значимые признаки даже при вариативности произношения, различиях в темпе и наличии акцентных искажений.
- *Рекуррентная сеть (RNN)* на базе LSTM или GRU отслеживает временные зависимости между акустическими фрагментами, что особенно важно при анализе русской речи с акцентом, где длительность фонем и структура слогов могут отличаться от речи носителей. Такая структура помогает учитывать контекст в пределах слова и фразы, а также сглаживать вариативность, связанную с интонацией и акцентом [10].

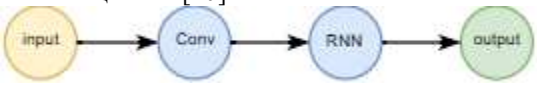

Рисунок 7 - архитектура CRNN

Благодаря сочетанию CNN и RNN, CRNN способны извлекать информативные акустические признаки и эффективно анализировать их во временной последовательности. В этом случае свёрточные слои выделяют акустические особенности речи, а рекуррентные слои моделируют временные зависимости между звуками, что позволяет распознавать слова и фразы даже при вариативности произношения.

III. Корпус и предобработка данных

Для исследования был использован корпус собственных записей русской речи, собранный с целью анализа того, как иностранный акцент влияет на качество распознавания. В создании корпуса участвовали дикторы с разным уровнем владения русским языком и разнообразными акцентными характеристиками. В выборку вошли участники, для которых русский язык не является родным, включая носителей европейских и азиатских акцентных особенностей. Такое сочетание обеспечило широкий спектр фонетических вариаций, встречающихся у иностранных говорящих.

Записи выполнялись на мобильные устройства дикторов в естественных бытовых условиях, что позволило сохранить акустическое разнообразие и приблизить данные к реальным сценариям использования. Полученные файлы были сохранены в формате .mp4, после чего аудиодорожки извлекались и конвертировались в формат .wav с частотой дискретизации 16 kHz.

Для повышения вариативности корпуса и моделирования дополнительных фонетических особенностей акцентированной речи применялись методы аугментации: изменение высоты тона, варьирование темпа и добавление фонового шума. Эти преобразования расширили объём тренировочного материала и повысили устойчивость моделей к акустическим отклонениям.

После подготовки аудиофайлы преобразовывались в мел-спектрограммы с использованием окна 25 мс, шага 10 мс и 80 мел-фильтров. На этапе получения спектрограмм проводилась нормализация, что снижало влияние разницы в громкости и улучшало стабильность обучения.

Все архитектуры обучались в единых условиях: каждая модель проходила 30 эпох обучения, использовался оптимизатор Adam, а в качестве функции потерь применялась CTC. Единый набор параметров позволил корректно сравнивать результаты и особенности работы различных моделей на акцентированной речи [11].

Для оценки качества распознавания использовались три метрики [12]:

- *Word Error Rate (WER)*, отражающая ошибки на уровне слов;
- *Character Error Rate (CER)*, позволяющая анализировать точность посимвольного распознавания;
- *Accuracy*, показывающая долю правильно распознанных символов.

Такой набор показателей обеспечивает комплексную оценку производительности моделей в условиях ограниченного и неоднородного корпуса акцентированной речи.

IV. Результаты исследования

Для сравнения архитектур были обучены семь

моделей: DNN, CNN, RNN, Stacked RNN, LSTM, Conformer и CRNN. Все архитектуры тестировались на одном и том же корпусе акцентированной русской речи и оценивались по метрикам WER, CER и Accuracy. Такой подход позволил выявить различия в их поведении при обработке нестандартного произношения и определить, какие модели лучше справляются с вариативностью акцентированного сигнала.

Таблица 1: Результаты обучения моделей на акцентированной русской речи

Модель	Loss	WER	CER	Accuracy	Время обучения
DNN	1.94	0.6	0.6	40.1%	24.22 мин
CNN	1.4	0.45	0.45	55.4%	25.9 мин
RNN	0.29	0.35	0.12	54.3%	41.25 мин
LSTM	0.23	0.33	0.12	56.4%	37.75 мин
Stacked RNN	0.57	0.53	0.2	37.3%	27.64 мин
Conformer	0.07	0.39	0.13	52.6%	27.77 мин
CRNN	0.08	0.57	0.23	36.9%	41.19 мин

Представленные значения показывают различия в поведении архитектур при работе с акцентированной русской речью. Наиболее низкие показатели WER и CER наблюдаются у LSTM и однослойной RNN, что говорит о лучшей способности рекуррентных моделей учитывать особенности временной структуры сигнала. CNN демонстрирует высокую Accuracy благодаря эффективному выделению локальных акустических признаков, однако по WER и CER уступает моделям, использующим рекуррентные механизмы. Stacked RNN показывает более высокие ошибки, что может быть связано с недостаточным объёмом данных для обучения глубокой рекуррентной архитектуры. Модели Conformer и CRNN занимают промежуточные позиции: самовнимание и сверточные блоки улучшают обработку спектральных особенностей, но ограниченность корпуса снижает их стабильность при распознавании акцентированных сегментов.

Помимо точностных метрик, был проведён сравнительный анализ скорости обучения. Время, затраченное на обучение одной модели при одинаковых условиях, заметно различается. Самыми быстрыми оказались DNN и CNN, для которых характерна относительно простая структура и отсутствие рекуррентных вычислений. RNN-модели (в том числе Stacked RNN и LSTM) обучались значительно дольше, что связано с последовательной обработкой временных шагов и высоким числом параметров. Наиболее ресурсоёмкими стали CRNN и классическая RNN, так как обе архитектуры совмещают как свёрточные, так и рекуррентные операции.

Несмотря на это, увеличение времени обучения не всегда приводит к снижению ошибок распознавания. Например, LSTM демонстрирует одно из лучших значений WER и CER при более длительном обучении, в то время как CRNN, даже при сопоставимой длительности эпох, показывает высокие WER и CER. Это позволяет отметить, что скорость обучения не является прямым показателем качества модели и требует анализа в сочетании с метриками точности.

Отдельно стоит отметить различия между рекуррентными и гибридными архитектурами. Модели с элементами самовнимания, такие как Conformer, требуют большего объёма данных для устойчивого обучения, особенно когда речь содержит акцентные искажения. В условиях ограниченного корпуса это приводит к повышенной чувствительности к вариативности произношения и ухудшению результатов по WER. CRNN, совмещающая сверточные и рекуррентные блоки, показывает более сбалансированное поведение, однако её точность также ограничена недостаточной длительностью и разнообразием входной последовательности.

Полученные показатели демонстрируют, что архитектуры, сохраняющие пошаговый контекст, лучше приспосабливаются к особенностям акцентированной речи. Это может быть связано с тем, что акцент влияет прежде всего на длительность фонем, последовательность артикуляционных переходов и стабильность мелодико-ритмических шаблонов. В таких условиях механизмы долгосрочной памяти оказываются более устойчивыми, тогда как модели, использующие самоориентированное внимание или глубокие сверточные преобразования, требуют более обширных корпусов для достижения стабильных результатов.

Помимо анализа различий между архитектурами при обучении на полном корпусе, было также проведено дополнительное исследование, направленное на оценку влияния объёма обучающих данных на качество распознавания. Для этого обучение повторялось на трёх вариантах выборки: 1/2 корпуса (3651 аудиозаписи), 2/3 корпуса (4819 аудиозаписи) и полный набор записей (7302 аудиозаписи). Такой подход позволил определить, насколько увеличение количества примеров влияет на устойчивость моделей к акцентным искажениям.

Таблица 2: Результаты модели LSTM при разных размерах обучающей выборки

Train fraction	Loss	WER	CER	Accuracy	Время обучения
50%	0.61	0.7	0.33	22.6%	22.77 мин
66%	0.34	0.53	0.22	36.8%	26.43 мин
100%	0.2	0.3	0.1	59.1%	38.73 мин

Поведение LSTM демонстрирует чёткую зависимость качества от размера выборки: при удвоении объёма данных WER снижается более чем в два раза, а Accuracy увеличивается почти в три раза. Это подтверждает

высокую чувствительность рекуррентных моделей к количеству обучающих примеров и необходимость использования достаточно больших корпусов для достижения низких ошибок распознавания.

Таблица 3: Результаты модели CRNN при разных размерах обучающей выборки

Train fraction	Loss	WER	CER	Accuracy	Время обучения
50%	0.17	0.78	0.39	19.5%	19.7 мин
66%	0.24	0.78	0.37	17.7%	24.43 мин
100%	0.18	0.61	0.25	30.5%	34.22 мин

CRNN демонстрирует менее гладкую динамику: качество на 66 % практически не улучшилось по сравнению с 50 %, что может быть связано с неоднородностью корпуса и вариативностью акцентированных записей. Однако при обучении на полном наборе данных наблюдается существенное улучшение, что подтверждает зависимость гибридных архитектур от объёма и разнообразия входного материала.

Таким образом, исследование зависимости точности от объёма данных указывает, что при разработке систем распознавания акцентированной речи качество модели может значительно улучшаться за счёт расширения корпуса. Особенно это касается архитектур, основанных на механизмах долгосрочной памяти и самовнимания, для которых размер выборки играет критическую роль.

Представленный анализ позволяет выявить архитектуры, демонстрирующие более надёжное поведение на небольшом и неоднородном датасете акцентированной русской речи. Итоги эксперимента формируют основу для выбора моделей, которые могут быть адаптированы и расширены в дальнейших исследованиях, направленных на улучшение распознавания речи носителей других языков.

V. Заключение

Проведённое исследование позволило оценить особенности работы различных нейросетевых архитектур при распознавании русской речи, произнесённой с иностранным акцентом. Эксперименты показали, что модели, использующие пошаговую обработку последовательности и внутреннее состояние (RNN и LSTM), достигают более низких значений WER и CER при работе с ограниченным набором данных. Корпус, использованный в исследовании, включал записи длиной в несколько часов и содержал речь дикторов с европейскими и азиатскими акцентными признаками. Этот объём, согласно проведённому анализу, является недостаточным для устойчивого обучения архитектур, требующих большого количества примеров, особенно Conformer и CRNN.

Дополнительные эксперименты с уменьшенными выборками (1/2 и 2/3 корпуса) показали, что снижение объёма данных наиболее заметно отражается на качестве гибридных моделей и моделей с механизмами самовнимания. При уменьшении обучающей выборки их показатели WER и CER возрастали сильнее всего. В то же время рекуррентные архитектуры демонстрировали более плавное снижение точности, что подтверждает их способность адаптироваться к вариативности акцентированного сигнала даже при ограниченном количестве данных.

Сверточные модели сохраняли стабильное значение Accuracy, что связано с эффективным выделением локальных спектральных признаков, но при уменьшении корпуса были менее устойчивы к вариативности ударения и темпа произнесения. Таким образом, выбор архитектуры напрямую зависит от доступного объёма данных: рекуррентные модели оказываются предпочтительными при работе с корпусами ограниченного размера, тогда как гибридные решения требуют заметно более обширной обучающей базы.

Полученные результаты могут быть использованы при создании систем автоматического распознавания русской речи для иностранных студентов, а также при дальнейшем расширении корпуса акцентированной речи и адаптации моделей под конкретные акцентные группы.

Библиография

[1] А.А. Волкова, Е.В. Дружинская Обзор моделей автоматического распознания акцентированной речи // Наука настоящего и будущего: материалы XII научно-практической конференции студентов, аспирантов и молодых учёных, 15–17 мая 2025 г. — СПб.: СПбГЭТУ «ЛЭТИ», 2025. — Т. 1. — С. 22–25.

[2] Deep Neural Network (DNN) Explained // Medium [Электронный ресурс]. - 2024. - URL: https://medium.com/@zomev/deep-neural-network-dnn-explained-0f7311a0e869 4 (дата обращения: 18.11.2025)

[3] Шишкин А.Г. Методы цифровой обработки и распознавания речи: монография / А.Г. Шишкин. - Москва: ИНФРА-М, 2024. - 347 с.

[4] S. Kostadinov Recurrent Neural Networks with Python Quick Start Guide. -Birmingham: Packt Publishing, 2018. - 122 p.

[5] Purwins H., Li B., Virtanen T., Schlüter J., Chang S., Sainath T. Deep Learning for Audio Signal Processing // IEEE Journal of Selected Topics of Signal Processing. - 2019. - Vol. 13. - No. 2. - p. 206–219.

[6] Тампель И.Б., Карпов А.А. Автоматическое распознавание речи. Учебное пособие. - СПб: Университет ИТМО, 2016. - 138 с.

[7] Patil S. Stacked RNNs in NLP // Artificial Intelligence in Plain English : [Электронный ресурс]. – 2023. – URL: https://python.plainenglish.io/stacked-mns-in-nlp-936e6eecf37a (дата обращения: 18.11.2025).

[8] Khan S., Naseer M., Hayat M., Zamir S.W., Khan F.S., Shah M. Transformers in Vision: A Survey // ACM Computing Surveys. - 2022. - Vol. 54. - No. 10. - p. 1–41.

[9] Как работают системы распознавания речи // Amvera [Электронный ресурс]. - 2022. - URL: https://amvera.ru/howasrwork (дата обращения: 18.11.2025).

[10] Varunanantharasa P. Building a Handwriting Recognition System with CRNN: A Beginner's Guide // Medium : [Электронный ресурс]. – 2025. – URL: https://medium.com/@pavitharan2020/building-a-handwriting-recognition-system-with-cmn-a-beginners-guide-58a51a46dd15 (дата обращения: 18.11.2025).

[11] Kheddar H., Hemis M., Himeur Y. Automatic Speech Recognition using Advanced Deep Learning Approaches: A survey // Information Fusion. – 2024. – Vol. 109. – No. 102422.

[12] Kamath U., Liu J., Whitaker J. Deep Learning for NLP and Speech Recognition. – Cham : Springer, 2019. – 621 p.

Comparison of neural network architectures for recognizing Russian speech with a foreign accent

A.A. Volkova, E.V. Druzhinskaya

Abstract—Automatic speech recognition systems are actively developing thanks to the widespread use of deep learning methods. However, when working with Russian speech pronounced by speakers of other languages, most algorithms encounter a decrease in accuracy. Accent specificity affects the duration of sounds, articulatory transitions, and prosodic characteristics, which makes it difficult to correctly identify acoustic features and subsequently transcribe them.

The scientific literature describes many approaches to speech processing, including deep, convolutional, and recurrent models, as well as hybrid architectures that use attention mechanisms. Each of them responds differently to pronunciation variability and accent intensity.

Taking these features into account, this work aims to study the behavior of various neural network architectures in recognizing Russian speech with a foreign accent and to analyze their results on a corpus of our own recordings. The evaluation is based on WER, CER, and Accuracy metrics, which allows us to identify models that demonstrate the greatest resistance to accent distortions and are capable of working with limited and heterogeneous data.

Keywords—speech recognition, accented speech, neural network architectures, speech processing, deep learning.

REFERENCES

[1] A.A. Volkova, E.V. Druzhinskaya Review of models of devices for recognizing accented speech // Science of the present and the future: materials of the XII scientific and practical conference of students, graduate students and young scientists, May 15–17, 2025. — St. Petersburg: ETU "LETI", 2025. — Vol. 1. — P. 22–25.

[2] Deep Neural Network (DNN) Explained // Medium [Online]. - 2024. - URL: https://medium.com/@zomev/deep-neural-network-dnn-explained-0f7311a0e8694 (accessed: 18.11.2025)

[3] Shishkin, A.G. Methods of Digital Speech Processing and Recognition: Monograph / A.G. Shishkin. - Moscow: INFRA-M, 2024. - 347 p..

[4] S. Kostadinov Recurrent Neural Networks with Python Quick Start Guide. -Birmingham: Packt Publishing, 2018. - 122 p.

[5] Purwins H., Li B., Virtanen T., Schlüter J., Chang S., Sainath T. Deep Learning for Audio Signal Processing // IEEE Journal of Selected Topics of Signal Processing. - 2019. - Vol. 13. - No. 2. - p. 206–219.

[6] Tampel I.B., Karpov A.A. Automatic speech recognition. Textbook. - St. Petersburg: ITMO University, 2016. - 138 p.

[7] Patil S. Stacked RNNs in NLP // Artificial Intelligence in Plain English : [Online]. – 2023. – URL: https://python.plainenglish.io/stacked-rnns-in-nlp-936e6eecf37a (accessed: 18.11.2025).

[8] Khan S., Naseer M., Hayat M., Zamir S.W., Khan F.S., Shah M. Transformers in Vision: A Survey // ACM Computing Surveys. - 2022. - Vol. 54. - No. 10. - p. 1–41.

[9] How speech recognition systems work // Amvera [Online]. - 2022. - URL: https://amvera.ru/howasrwork (accessed: 18.11.2025).

[10] Varunanantharasa P. Building a Handwriting Recognition System with CRNN: A Beginner's Guide // Medium : [Online]. – 2025. – URL: https://medium.com/@pavitharan2020/building-a-handwriting-recognition-system-with-crnn-a-beginners-guide-58a51a46dd15 (accessed: 18.11.2025).

[11] Kheddar H., Hemis M., Himeur Y. Automatic Speech Recognition using Advanced Deep Learning Approaches: A survey // Information Fusion. – 2024. – Vol. 109. – No. 102422.

[12] Kamath U., Liu J., Whitaker J. Deep Learning for NLP and Speech Recognition. – Cham : Springer, 2019. – 621 p.