DOI: 10.15514/ISPRAS-2025-37(1)-9

Глубокое обучение в задаче разработки системы автоматической транскрипции

О.В. Гончарова, ORCID: 0000-0003-1044-6244 <goncharovaov@pgu.ru>

Пятигорский государственный университет,
Россия, 357532, г. Пятигорск, Ставропольский край, пр. Калинина, 9,
Российский университет дружбы народов имени Патриса Лумумбы,
Россия, 117198, г. Москва, ул. Миклухо-Маклая, 6,
Институт системного программирования им. В.П. Иванникова РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

Аннотация. В статье представлена архитектура глубокой нейронной сети для автоматического распознавания фонем в речевом сигнале. Предложенная модель использует комбинацию сверточных и рекуррентных слоев, а также механизм внимания, обогащенный референсными значениями формант гласных фонем. Это позволяет эффективно извлекать локальные и глобальные акустические признаки, необходимые для точного распознавания последовательностей фонем. Особое внимание уделяется проблеме несбалансированности частоты фонем в обучающем наборе данных и способам ее преодоления, таким как аугментация данных и применение взвешенной функции потерь. Представленные результаты демонстрируют работоспособность предложенного подхода, однако указывают на необходимость дальнейшего совершенствования модели для достижения более высоких показателей точности и полноты в задаче распознавания речи.

Ключевые слова: Автоматическое распознавание речи; фонетическая транскрипция; глубокие нейронные сети; форманты.

Для цитирования: Гончарова О.В. Применение глубокого обучения для разработки системы автоматической транскрипции. Труды ИСП РАН, том 37, вып. 1, 2025 г., стр. 145–158. DOI: 10.15514/ISPRAS–2025–37(1)–9.

Благодарности: Исследование подготовлено в рамках гранта РНФ № 23-28-10124 «Квантитативно-статистическая модель анализа эмоционально-маркированной коммуникации в условиях межэтнических взаимодействий в регионе Кавказские Минеральные Воды».

Deep Learning for an Automatic Transcription System Development

O.V. Goncharova, ORCID: 0000-0003-1044-6244 <goncharovaov@pgu.ru>
Pyatigorsk State University,
9, Kalinin Ave., Pyatigorsk, Stavropol Rg., 357532, Russia,
The Patrice Lumumba Peoples' Friendship University of Russia,
6, Miklukho-Maklaya st., Moscow, 117198, Russia,
Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

Abstract. This paper presents a deep neural network architecture for automatic phoneme recognition in speech signals. The proposed model combines convolutional and recurrent layers, as well as an attention mechanism enriched with reference values of vowel formant frequencies. This allows the model to effectively extract local and global acoustic features necessary for accurate phoneme sequence recognition. Particular attention is paid to the problem of imbalanced phoneme frequency in the training dataset and ways to overcome it, such as data augmentation and the use of a weighted loss function. The reported results demonstrate the viability of the proposed approach, but also indicate the need for further model refinement to achieve higher accuracy and recall in the speech recognition task.

Keywords: Automatic speech recognition; phonetic transcription; deep neural networks; formants.

For citation: Goncharova O.V. Deep Learning for the Development of an Automatic Transcription System. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 1, 2025. pp. 145-158 (in Russian). DOI: 10.15514/ISPRAS-2025-37(1)-9.

Acknowledgements. The study was prepared within the framework of the Russian National Science Foundation Grant No. 23-28-10124 "Quantitative statistical model for the analysis of emotionally marked communication in the context of interethnic interactions in the Caucasian Mineral Waters region".

1. Введение

Современные технологии обработки естественного языка (Natural Language Processing, NLP) и распознавания речи являются фундаментальными компонентами в развитии систем взаимодействия человека и машины. С ростом объёмов аудиоинформации и возрастающей потребностью в её автоматической обработке возникает необходимость в высокоточных моделях, способных эффективно распознавать и интерпретировать звуковые данные.

Фонетические транскрипции являются одной из ключевых единиц в области обработки речи и повсеместно используются в лингвистических исследованиях как для научного анализа, так и для проверки гипотез [1-2]. Однако проблемы, связанные с их получением – значительные временные и финансовые затраты, а также ограниченная точность – становятся особенно актуальными в контексте обработки больших объёмов данных в современных речевых технологиях.

В связи с указанными трудностями исследователи стремятся автоматизировать процесс фонетической транскрипции посредством алгоритмов распознавания речи и разработка инструментов, способных автоматически выполнять разметку аудиоданных на фонетическом уровне, приобретает особую значимость. Такие средства не только ускоряют процессы лингвистического анализа и повышают качество систем распознавания речи, но и способствуют более глубокому пониманию фонетических особенностей языковых данных.

Автоматическая фонетическая транскрипция играет существенную роль не только в создании новых транскрипций, но и в проверке точности уже существующих. Она позволяет обнаруживать расхождения между реальным звучанием и записанной транскрипцией при аннотировании аудиозаписей. Вопрос оценки качества транскрипций также является актуальным: даже сделанные вручную фонетические транскрипции могут содержать

ошибки, поэтому их точность необходимо оценивать перед использованием. Независимо от того, были ли транскрипции получены автоматически или созданы вручную, они служат базой для последующей обработки, включая лингвистический анализ и обучение систем автоматического распознавания речи (ASR). Поскольку транскрипции рассматриваются как отображение или измерение речевого сигнала, важно определить степень их соответствия стандартам качества, надёжности и точности, требуемым от любых форм измерений.

В современном контексте для автоматизации процесса фонетической транскрипции речи существуют два разных подхода: независимая от текста сегментация фонем и выравнивание фонем по речи (или принудительное выравнивание). В настоящее время существует несколько инструментов для принудительного выравнивания [3-5], основанных преимущественно на классической системе HMM (Hidden Markov model – скрытая марковская модель), построенной на базе Kaldi [6] или HTK toolkit [7]. К наиболее распространённым относятся программы Forced Alignment and Vowel Extraction [5], Language, Brain and Behavior Corpus Analysis Tool [8], Montreal Forced Aligner [4] и Munich Automatic Segmentation System [3]. Например, работа инструмента MAUS (Мюнхенская автоматическая сегментация) организована следующим образом: пользователи загружают в онлайн-сервис орфографические транскрипции и аудиофайлы высказываний. Затем инструмент Balloon [9] преобразует текст в фонемы с использованием алгоритма преобразования графемы в фонему и словаря исключений. Далее на основе фонологических правил формируются возможные варианты произношения, в результате чего получается направленный ациклический график. Этот график представляет множество априорных статистически взвешенных гипотетических вариантов произношения высказывания, после чего система ASR, использующая скрытые марковские модели (HMM), определяет наиболее вероятный вариант произношения для каждого слова и устанавливает границы сегментов [3].

Однако несмотря на широкое распространение и полезность систем принудительного фонетического выравнивания, они имеют ряд существенных недостатков. Прежде всего, их эффективность напрямую зависит от наличия точных орфографических транскрипций, которые не всегда доступны или могут содержать ошибки, что может быть особенно критичным для малоресурсных языков или диалектов, где корпуса с орфографическими транскрипциями ограничены или вовсе отсутствуют. Кроме того, такие системы часто не учитывают вариативность спонтанной речи, включая диалектные особенности, редукции, ассимиляции и другие фонетические процессы, не отраженные в орфографии. Это приводит к снижению точности выравнивания и ограничивает возможности анализа реальных языковых данных. Работ по независимой от текста сегментация фонем значительно меньше и, как правило, предложенные инструменты менее практичны [10]. В свете указанных ограничений возникает необходимость в разработке альтернативных методов фонетической разметки, не зависящих от орфографического текста. Создание нейросетевых моделей, способных напрямую извлекать фонетическую информацию из аудиосигнала, представляется перспективным направлением. Такие модели, обученные на больших объемах данных, учитывающих разнообразие фонетических реализаций, позволят получать более точную и гибкую фонетическую разметку, применимую как к распространенным, так и к малоресурсным языкам, а также к различным речевым стилям и условиям записи.

В рамках настоящего исследования мы сосредоточимся на первом подходе, т.е., независимой от текста сегментация фонем. Предполагается, что описываемая архитектура нейросети будет полезна для автоматизации процесса фонетического анализа и проверки корректности существующих транскрипций, а также для создания обширных и точных корпусов речи для последующих исследований и практических приложений.

2. Методика проведения исследования и предварительная обработка данных

Для эффективного обучения моделей автоматического распознавания транскрипционных знаков нами была проведена комплексная работа по сбору и подготовке наборов аудиозаписей. Материалом являются аудиозаписи, собранные в различных условиях и объединённые в единый корпус для целей настоящего исследования. Общий объём данных составляет 4890 записей в формате .wav и соответствующих им аннотаций в формате TextGrid. Корпус сформирован на основе трёх ключевых групп записей:

1. Спонтанная речь в естественных условиях: включает 5 часов спонтанной речи 32 участников, полученной из ранее проведённых исследований [11]. Данная группа представляет фонетические варианты русского языка, зафиксированные в естественной коммуникации. Участники общались в непринуждённой обстановке, что позволит модели анализировать особенности спонтанной речи.
2. Лабораторные записи билингвов с эмоциональной окраской: содержит 12 часов речи 62 участников, собранных в рамках настоящего проекта. Исследование сфокусировано на изучении русской речи билингвов с различной эмоциональной экспрессией. Участники – билингвы, свободно владеющие русским языком; им предлагалось произносить фразы, отражающие различные эмоциональные состояния, в частности гнев и радость.
3. Данные из открытых источников платформы Lingvodoc: включает 2 часа лабораторных записей 12 участников [12]. В этой группе представлены преимущественно трёхкратные повторения стословника Сводеша, выполненные носителями урало-алтайских языков. Записи проводились в контролируемых условиях с целью обеспечения высокого качества и сопоставимости данных.

Все аудиозаписи сопровождены аннотациями в формате TextGrid, где границы звуков были вручную размечены и проверены тремя независимыми аудиторами-лингвистами с использованием программного обеспечения Praat [13]. Записи осуществлялись с применением высококачественного аудиооборудования при частоте дискретизации 44,1 кГц в условиях, минимизирующих влияние посторонних шумов.

При формировании набора данных особое внимание уделялось обеспечению максимального разнообразия по ряду ключевых параметров: полу, возрасту, диалектной принадлежности и эмоциональному состоянию говорящих. Такой методологический подход позволил создать репрезентативную выборку, повышающую обобщающую способность обучаемых моделей и расширяющую их применимость к широкому спектру речевых особенностей и вариативности естественной речи.

Таким образом, сформированный корпус данных предоставляет надёжную основу для обучения и оценки моделей автоматического распознавания транскрипционных знаков, способствуя развитию более точных и универсальных систем обработки речи.

В процессе анализа и предварительной обработки собранных транскрипций была выявлено неоднородность используемых систем транскрипции и символов. Данный фактор создавал значительные сложности для последующего обучения моделей, поскольку использование разнородных транскрипционных систем приводит к несогласованности данных и потенциально снижает эффективность моделей. Для преодоления этой проблемы нами был осуществлен ряд преобразований файлов аннотаций, направленных на унификацию и стандартизацию транскрипционных данных:

- Стандартизация транскрипций: для обеспечения консистентности данных нами был разработан скрипт на языке Python, предназначенный для автоматизации процесса конвертации и обработки транскрипционных данных. Скрипт [14] осуществляет преобразование существующих транскрипционных символов в формат

Международного фонетического алфавита (IPA), что позволяет унифицировать данные и облегчить их последующий анализ.

- Разработка алгоритмов для преобразования файлов с аннотациями: нами был создан подробный словарь соответствий (mapping), где ключами являются исходные символы, а значениями – соответствующие им символы IPA. Если символ интервала присутствовал в словаре, скрипт заменял оригинальный символ на его эквивалент в IPA. Процесс сопровождался проверкой на необходимость изменения (то есть, если исходный и преобразованный символы не совпадают) и ведением статистики о количестве модифицированных интервалов (см. табл. 1).

Табл. 1. Фрагмент словаря соответствий символов не входящих в IPA.
Table 1. A fragment of the dictionary of matches of characters not included in the IPA.

Значение в исходном файле TextGrid	Значение после работы скрипта
'r'	'g'
"\\ng"	"ŋ"
"tt͡ʃ"	"t͡ʃ"

- Обработка нераспознанных и неоднозначных символов: в случаях, когда символ интервала не был найден в словаре соответствий, скрипт проверял его наличие в списке (unassigned_symbols), который содержал символы, не поддающиеся однозначному сопоставлению или отсутствующие в стандарте IPA. Если символ был обнаружен в этом списке, интервал считался проблемным и ему присваивалось значение “unknown”. Данное решение позволило избежать включения неверных или неоднозначных данных в итоговый набор транскрипций, сохраняя при этом возможность дальнейшего анализа этих случаев отдельно (см. табл. 2).

Табл. 2. Фрагмент словаря «unassigned_symbols» неизвестных символов и символов не входящих в IPA.
Table 2. A fragment of the «unassigned_symbols» dictionary for unknown characters and characters not included in the IPA.

Значение в исходном файле TextGrid	Значение после работы скрипта
"u\\:f"	unknown
"x̆͡ʲç"	unknown
"čč"	unknown
"c̆"	unknown
"ɦ͡нᵝ"	unknown
"œ̬"	unknown
"β̬"	unknown
"ú"	unknown

- Исключение «unknown» символов при обучении: все транскрипционные знаки, помеченные как «unknown», были замаскированы или исключены из обучающей выборки. Данное решение было принято для предотвращения внесения шумовых данных и обеспечения чистоты обучающего набора, что способствует повышению производительности и точности моделей распознавания (см. табл. 3).

Далее был сформирован полный набор уникальных фонем, присутствующих в транскрипциях аудиозаписей. Для этого из каждой транскрипции извлекался список фонем, при этом были исключены все неопознанные или неизвестные символы, замененные на «unknown». Полученный набор фонем был объединен и приведен к множеству уникальных значений, что позволило определить полный список фонем, которые необходимо распознавать модели, исключая шумовые и некорректные данные.

На основе извлеченного набора был создан упорядоченный список фонем. Каждой фонеме был присвоен уникальный числовой идентификатор, формируя таким образом словарь

фонем. Дополнительно в словарь был добавлен специальный символ «|», предназначенный для представления пауз или возможных пропущенных в ходе обработки символов. Данный словарь служил для преобразования последовательностей фонем из транскрипций в числовые последовательности, которые были использованы при обучении нейронной сети.

Табл. 3. Фрагмент обработанного набора данных с заменой неизвестных символов.
Table 3. A fragment of the processed dataset with the replacement of unknown characters.

Название файла	Транскрипция
382.wav	j unknown unknown j
72.wav	d r ɤ
120181.wav	s unknown l unknown s ɛ n ø k s h ɛ unknown p unknown y j ɘ s
1953.wav	ʃ d ɐ r unknown
745.wav	p unknown l z unknown q
66.wav	t unknown n unknown d l
355.wav	p ɜ m unknown
2255.wav	e l ɨ o
2241.wav	k unknown t
1748.wav	s v unknown t j unknown s v a t j unknown s v unknown t j œ
2094.wav	k ɤ n d unknown k
221.wav	p y j p y j p y j
547.wav	ø t unknown unknown unknown ɔ
2080.wav	n ø m
631.wav	ɤ l d œ
1210.wav	j e m b ɘ l j e m b œ l j e m b ɘ l
139.wav	ø d ɘ ø ɛ
2862.wav	j a v ɜ p unknown m ɑ l ʌ k ʌ z unknown f t r unknown

Ключевым этапом подготовки данных явилась токенизация, включающая:

- Загрузку и предварительную обработку аудиоданных: каждая аудиозапись загружается и приводится к единой частоте дискретизации – 16 кГц. Это обеспечивает стандартизацию аудиоданных и повышение качества извлеченных признаков.
- Обработку транскрипций: транскрипции разбиваются на отдельные фонемы. Все вхождения слова «unknown» маскируются, чтобы исключить неопознанные символы из дальнейшего анализа.
- Кодирование фонем: каждая фонема из транскрипции преобразуется в соответствующий числовой индекс с помощью ранее созданного словаря фонем. Если фонема по каким-либо причинам отсутствует в словаре, ей присваивается индекс специального символа «|», что позволяет обработать все возможные случаи без возникновения ошибок.

- Извлечение признаков из аудиоданных: используя инициализированный аудиопроцессор, аудиоданные преобразуются в тензоры входных значений, подходящие для подачи на вход нейронной сети. Процессор производит необходимую нормализацию и, при необходимости, выравнивание по длине (padding).
- Формирование меток: закодированные числовые последовательности фонем используются в качестве меток (labels) для обучения модели – это связывает каждую аудиозапись с соответствующей последовательностью фонемных индексов.
- Объединение данных: обработанные аудиоданные и соответствующие им метки объединяются в единый набор данных, готовый для обучения нейронной сети.

Процесс токенизации и предварительной обработки данных был применен ко всему набору данных, таким образом, каждый образец в наборе данных прошел через описанные этапы преобразования, что обеспечивает единообразие и согласованность данных.

3 Описание признаков

В настоящем исследовании в качестве стандартных акустических признаков использовались мел-частотные кепстральные коэффициенты (MFCC) и мел-спектрограммы (Melspec). MFCC представляют собой спектральные характеристики, отражающие энергетическое распределение сигнала по частотам на нелинейной мел-шкале, которая соответствует восприятию частот человеческим слухом.

Процесс извлечения MFCC включает следующие этапы:

1. Преобразование в частотную область: Исходный речевой сигнал разбивается на короткие временные отрезки (фреймы), к которым применяется быстрое преобразование Фурье (FFT) для перехода в частотную область.
2. Преобразование спектра в мел-шкалу: Полученные спектры мощности проходят через банк мел-фильтров, что позволяет акцентировать частоты, значимые для человеческого восприятия.
3. Логарифмирование и обратное преобразование: Логарифм спектра в мел-шкале вычисляется для сжатия динамического диапазона, после чего применяется обратное дискретное косинусное преобразование (IDCT) для получения MFCC.

Для учета динамических характеристик речи были вычислены производные первого и второго порядка от MFCC:

- Δ (дельта) коэффициенты: представляют скорость изменения MFCC во времени, что позволяет моделировать переходные процессы между фонемами.
- ΔΔ (дельта-дельта) коэффициенты: отражают ускорение изменений MFCC, захватывая более сложные динамические аспекты речевого сигнала.

Форманты являются резонансными частотами голосового тракта, возникающими при артикуляции звуков речи. Особенно важны первые две форманты, F1 и F2, для различения гласных звуков:

- F1: соответствует вертикальному положению языка (степени открытости гласного звука).
- F2: связан с горизонтальным положением языка (передний, средний или отодвинутый назад ряд гласного звука).

Извлеченные формантные частоты сравнивались с типичными значениями для различных гласных звуков на основе фонетических исследований [11]. Для каждого гласного вычислялась абсолютная разница между эталонными формантными частотами и измеренными значениями F1 и F2.

Среднеквадратическое значение энергии (RMS) использовалось для оценки средней амплитуды сигнала в каждом фрейме, что характеризует громкость звука.

Все упомянутые признаки объединялись в единый вектор для каждого временного фрейма. В результате каждый фрейм представлен набором характеристик, отражающих его спектральные, динамические, артикуляционные и энергетические свойства.

4 Архитектура нейронной сети

Наша цель заключалась в создании системы, способной с высокой точностью распознавать фонемы в потоке речи, используя современные методы глубокого обучения и обработки сигналов. В последние годы модели на основе глубоких нейронных сетей продемонстрировали значительный прогресс в задачах автоматического распознавания речи [15, 16], а использование трансформерных архитектур улучшило качество обработки последовательностей [17].

В основе модели лежит позиционное кодирование, которое учитывает порядок звуков в речи. Поскольку последовательность элементов имеет решающее значение для понимания смысла высказывания, мы добавили информацию о позиции каждого звука в последовательности. Такой подход соответствует практике применения позиционного кодирования в трансформерных моделях для учета порядка элементов [17], что позволяет модели различать одинаковые звуки в разных контекстах и учитывать синтагматические связи между фонемами, немаловажно, что данный подход ранее был успешно продемонстрирован в работах по обработке естественного языка [18].

При обучении модели мы столкнулись с неравномерным распределением фонем и разной значимостью ошибок в их распознавании. Для решения проблемы мы разработали пользовательскую функцию потерь с весами и маскированием. Подобные методы взвешивания классов также ранее применялись для борьбы с дисбалансом данных в задачах классификации [19, 16]. Разработанная функция позволяет назначать разные веса разным классам фонем, что обеспечивает фокусировку модели на более редких или критически важных звуках. Маскирование используется для игнорирования нерелевантных или отсутствующих данных в последовательности, повышая устойчивость модели к шумам и паузам [20].

На вход модель принимает последовательности акустических признаков, включающих как общие характеристики звука, так и специфические акустические параметры (см. раздел 3).

Для извлечения локальных паттернов в акустическом сигнале используются свёрточные слои (Conv1D), предназначенные для выявления значений частот или изменения амплитуды, характерные для конкретных фонем [21]. Следом за ними идут рекуррентные слои на основе архитектуры "Long short-term memory" (LSTM), которые позволяют модели сохранять и использовать информацию о предыдущих элементах последовательности, что особенно важно для учёта контекста и последовательности звуков в речи [22].

Далее в архитектуре применяются двунаправленные рекуррентные слои (Bidirectional LSTM), позволяя модели учитывать как предыдущий, так и последующий контекст при распознавании каждой фонемы. Следующие далее механизмы внимания (Attention) интегрированы в модель для того, чтобы она могла выделять наиболее значимые части последовательности при предсказании текущей фонемы [23]. В качестве дополнительного слоя внимания мы добавили референсные значения формант гласных фонем. Данные значения представляют собой эталонные частоты первых двух формант (F1 и F2) для каждой гласной, что позволяет модели более точно сопоставлять входные акустические признаки с соответствующими фонемами (см. табл. 4) [24, 11].

Представленные в табл. 4 значения были рассчитаны в диапазоне ±100 Гц для учёта индивидуальных вариаций и особенностей произношения. Предполагалось, что интеграция

референсных значений формант в механизм внимания позволит модели более точно выделять гласные фонемы, даже при наличии акцентов или фонетических искажений [25].

Чтобы предотвратить переобучение и улучшить генерализацию модели, мы применили слои нормализации и регуляризации [26], что особенно важно при работе с аудиозаписями, содержащими различные варианты произнесения.

В выходных слоях модель преобразует внутренние представления в вероятностное распределение по классам фонем. На основе этого распределения она принимает решение о конкретной фонеме в заданной позиции последовательности, обеспечивая точное и последовательное распознавание.

Табл. 4. Референсные значения гласных фонем.
Table 4. Reference values of vowel phonemes.

i	240, 2400
e	390, 2300
ɛ	610, 1900
a	850, 1610
ɑ	750, 940
ɔ	500, 700
o	360, 640
u	250, 595
y	235, 2100
ø	370, 1900
œ	585, 1710
ɶ	820, 1530
ɒ	700, 760
ʌ	600, 1170
ɤ	460, 1310
ɯ	300, 1390
ɨ	300, 1660
ʉ	320, 1390
ɘ	400, 1720
ɵ	430, 1390
ɜ	550, 1640
ɞ	560, 1330

Общая концепция работы модели заключается в том, что она принимает на вход аудиофайл речи, который предварительно преобразуется в последовательность акустических признаков с учётом формантных характеристик. Позиционное кодирование добавляет информацию о расположении каждого звука. Затем, через комбинацию свёрточных и рекуррентных слоёв, модель извлекает как локальные, так и глобальные зависимости в данных. Механизм внимания, обогащённый референсными значениями гласных, позволяет сконцентрироваться на наиболее значимых признаках для текущего предсказания. На выходе модель генерирует последовательность фонем, осуществляя тем самым фонетическую транскрипцию исходного аудиосигнала.

Поскольку в данных могут присутствовать позиции, которые не следует учитывать при вычислении ошибки (например, заполнители или специальные токены), была разработана взвешенная маскированная функция потерь WeightedMaskedLoss, основанная на категориальной перекрестной энтропии и включающая в себя механизмы маскирования и взвешивания классов. Принцип расчета функции потерь:

1. Маскирование незначимых позиций: вычисляется маска mask, представляющая собой тензор, где элементы равны 1, если соответствующая метка верного значения

(y_true) не равна -1 (то есть значима для обучения), и 0 в противном случае. Это позволяет исключить из расчета ошибки те позиции, которые не несут полезной информации.

2. Преобразование меток: исходные метки y_true преобразуются в целочисленный формат y_true_int. Затем с помощью маски формируется новый тензор меток y_true_fixed, где незначимые позиции заменяются на нули. Это необходимо для корректного вычисления перекрестной энтропии.
3. Вычисление базовой ошибки: с помощью функции sparse_categorical_crossentropy рассчитывается базовая ошибка между предсказанными значениями y_pred и преобразованными метками y_true_fixed.
4. Применение весовых коэффициентов классов: для компенсации дисбаланса классов используется вектор весовых коэффициентов weights. Путем создания one-hot представления меток y_true_one_hot и последующего умножения на веса получается тензор class_weights, содержащий веса для каждого примера. Затем ошибка умножается на соответствующие веса классов.
5. Применение маски к ошибке: ошибка умножается на маску mask, чтобы исключить вклад незначимых позиций.
6. Нормализация ошибки: итоговая ошибка рассчитывается как сумма взвешенных ошибок, деленная на сумму элементов маски. Это обеспечивает корректное усреднение ошибки только по значимым позициям.

5 Результаты

Оценка метрик производительности разработанной модели свидетельствует о достижении следующих показателей: Точность (Precision) – 77%, Полнота (Recall) – 65%, F1-мера – 70,50%, и общая Точность классификации (Accuracy) – 65,77%. Для более глубокого понимания классификационных способностей модели был использована матрица несоответствий (см. рис. 1). Матрица несоответствий (confusion matrix) является широко используемым инструментом для оценки классификационных моделей, поскольку сопоставляет фактические классы данных с предсказанными моделью, тем самым выявляя типы и частоты ошибок [27].

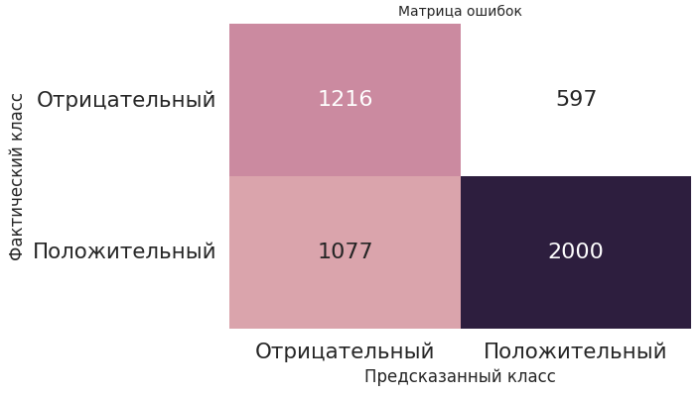

Рис. 1. Матрица несоответствий модели автоматической транскрипции.
Fig. 1. The confusion matrix of the automatic transcription model.

Анализируя матрицу несоответствий, было отмечено, что количество правильно предсказанных положительных экземпляров (истинно положительные, когда класс 1 правильно предсказан как класс 1) составляет 2000, количество правильно предсказанных

отрицательных экземпляров (истинно отрицательные, когда класс 0 правильно предсказан как класс 0) – 1216. Однако имеются значительные ошибки классификации: 1077 случаев, когда положительные экземпляры были неверно предсказаны как отрицательные (ложноотрицательные), и 597 случаев, когда отрицательные экземпляры были неверно предсказаны как положительные (ложноположительные).

Модель демонстрирует более высокую точность для класса 1 (положительный класс) – 77% по сравнению с точностью 65% для класса 0 (отрицательный класс). Одним из потенциальных факторов, влияющих на производительность модели, является несбалансированность частоты уникальных фонем в наборе данных. Анализ распределения фонем выявляет существенные диспропорции в их встречаемости (см. табл. 5):

- Фонемы с высокой частотой: 'l' (2153 вхождения), 't' (1942), 'n' (1936), 'm' (1853), 'k' (1807), 'p' (1673).
- Фонемы с низкой частотой: 'ʧ' (115 вхождений), 'ɵ' (191), 'f' (163), 'ɒ' (166), 'u' (172), 'w' (346).

Табл. 5. Частота встречаемости уникальных фонем в наборе данных.
Table 5. The frequency of occurrence of unique phonemes in the dataset.

m: 1853	v: 997
j: 1687	ʌ: 1050
ɤ: 999	q: 680
d: 1293	z: 407
r: 1449	ʒ: 751
p: 1673	i: 622
l: 2153	e: 1075
ø: 1182	a: 947
k: 1807	w: 346
ɘ: 1008	o: 513
s: 981	ɨ: 1138
n: 1936	ʧ: 115
y: 491	ɵ: 191
ɛ: 1094	b: 241
h: 377	ɯ: 268
ʃ: 1023	f: 163
ɞ: 962	ɑ: 280
ɔ: 1092	ʉ: 192
x: 518	ɒ: 166
œ: 1211	u: 172
ɶ: 740	t: 1942
unknown: 3813 (не учитывается при обучении)	

Данная несбалансированность указывает на то, что некоторые фонемы представлены чрезмерно, тогда как другие – недостаточно. Например, фонема 'l' встречается почти в 19 раз чаще, чем 'ʧ'. Такие диспропорции могут привести к смещённому процессу обучения, при котором модель становится более способной распознавать шаблоны, связанные с частыми фонемами, в то время как ей не хватает данных для обучения на редких фонемах.

Неравномерное распределение фонем может существенно повлиять на способность модели обобщать результаты по всем классам. В задачах классификации модели, обученные на несбалансированных наборах данных, склонны отдавать предпочтение более

многочисленному классу, что приводит к более высоким уровням ошибок для малочисленных классов [28]. В контексте задач классификации фонем это означает, что модель может неэффективно изучать характеристики менее частых фонем, что приводит к увеличению количества ошибок, когда эти фонемы присутствуют во входных данных.

Например, модель может правильно классифицировать слова, содержащие высокочастотные фонемы, такие как 'l', 't' или 'n', благодаря обилию обучающих примеров. Напротив, она может ошибочно классифицировать слова, содержащие низкочастотные фонемы, такие как 'tʃ' или 'f', поскольку она недостаточно изучила их связанные шаблоны. Это может способствовать увеличению количества ложноотрицательных и ложноположительных результатов, как это наблюдается в матрице неточностей.

4 Заключение

В результате анализа был сделан вывод о том, что невысокая точность модели в 65,77% отчасти является следствием несбалансированности частоты фонем в наборе данных. Чрезмерное представление некоторых фонем, таких как 'l', 't' и 'n', и недостаточное представление других, таких как 'tʃ', 'f' и 'u', создают условия обучения, при которых модель не учится одинаково на всех примерах. Чтобы смягчить влияние несбалансированности фонем на производительность модели, можно применять несколько стратегий:

1. Аугментация Данных: Увеличение количества экземпляров, содержащих редкие фонемы, с помощью методов аугментации данных может помочь сбалансировать набор данных и предоставить модели больше примеров для обучения [1].
2. Методы балансировки классов: Применение методов, например, с понижением количества экземпляров большинства (under-sampling) может скорректировать распределение классов в наборе данных, что потенциально улучшит способность модели к обобщению [24].
3. Инженерия Признаков: Внедрение методов взвешивания фонем или внедрения обученных векторов признаков, которые учитывают частоту фонем, может помочь модели уделять больше внимания малочисленным фонемам во время обучения.

Таким образом, устранение несбалансированности частоты фонем является критически важным для повышения производительности модели. Путём реализации стратегий по балансировке набора данных и соответствующей настройке алгоритмов обучения возможно улучшить способность модели точно классифицировать как частые, так и редкие фонемы, тем самым увеличивая общую точность. В целом, полученные результаты подтверждают работоспособность предложенного подхода, однако указывают на необходимость дальнейшего совершенствования модели и метода обучения для достижения более высоких показателей точности и полноты в задаче распознавания речи. В дальнейшем планируется взять больший объем однородного материала на одном языке, а затем на близкородственных языках.

Список литературы / References

[1]. Shorten, C., Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning // Journal of Big Data, 6(1):60, 2019. Доступно по ссылке: https://www.researchgate.net/publication/334279066_A_survey_on_Image_Data_Augmentation_for_Deep_Learning (Дата обращения 23.01.2025).

[2]. Cucchiarini, C., 1993. Phonetic transcription: a methodological and empirical study. Ph.D. thesis, University of Nijmegen, Nijmegen, The Netherlands. Доступно по ссылке: https://repository.ubn.ru.nl/bitstream/handle/2066/145701/mmubn000001_170795853.pdf (Дата обращения 23.01.2025).

[3]. Kisler T., Schiel F., Sloetjes, H. Signal processing via web services: the use case WebMAUS. // Digital Humanities Conference 2012. 2012. pp. 30-34. Доступно по ссылке:

https://www.researchgate.net/publication/248390251_Signal_processing_via_web_services_the_use_case_WebMAUS (Дата обращения 23.01.2025).
[4]. McAuliffe M., Socolof M., Mihuc S., Wagner M., Sonderegger M., Montreal forced aligner: Trainable text-speech alignment using Kaldi // Proc. Interspeech, vol. 2017. 2017. pp. 498– 502.
[5]. Rosenfelder I., Fruehwald J., Evanini K., Yuan, J. FAVE (forced alignment and vowel extraction) program suite. 2011. Доступно по ссылке: http://fave.ling.upenn.edu (Дата обращения 23.01.2025).
[6]. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlíček P., Qian Y., Schwarz P., Silovský J., & Stemmer G., Vesel K. The Kaldi Speech Recognition Toolkit // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011. Доступно по ссылке: https://www.danielpovey.com/files/2011_asru_kaldi.pdf (Дата обращения 23.01.2025).
[7]. Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P., The HTK book // Cambridge university engineering department, vol. 3, no. 175, pp. 12. 2002. Доступно по ссылке: https://www.danielpovey.com/files/htkbook.pdf (Дата обращения 23.01.2025).
[8]. Fromont R., Hay J. LaBB-CAT: an Annotation Store // Proceedings of the Australasian Language Technology Association Workshop 2012, 2012. pp. 113–117. Доступно по ссылке: https://aclanthology.org/U12-1015.pdf (Дата обращения 23.01.2025).
[9]. Uwe R. PermA and Balloon: Tools for string alignment and text processing // paper no. 346. 2012. doi: 10.21437/Interspeech.2012-509 (Дата обращения 23.01.2025).
[10]. Teytaut Y., Roebel A. Phoneme-to-Audio Alignment with Recurrent Neural Networks for Speaking and Singing Voice // Proceedings of Interspeech 2021, International Speech Communication Association, Aug 2021, Brno, Czech Republic. pp.61-65, 10.21437/interspeech.2021-1676. hal-03552964 Доступно по ссылке: https://hal.science/hal-03552964/file/1676anav.pdf (Дата обращения 23.01.2025).
[11]. Гончарова О.В. Артикуляционно-акустические характеристики безударных и ударных гласных на месте орфографического ‘а’ в речи носителей разных фоновариантов русского языка // Филологические науки. Вопросы теории и практики. 2024. Том 17. Выпуск 5. 2024. Volume 17. С. 1661-1668. Доступно по ссылке: https://philology-journal.ru/article/phil20240240/fulltext (Дата обращения 23.01.2025).
[12]. Веб-сайт https://lingvodoc.ispras.ru/dictionaries_all (Дата обращения 23.01.2025).
[13]. Boersma P., Weenink D. PRAAT: Doing phonetics by computer. 2024. Доступно по ссылке: https://www.fon.hum.uva.nl/praat/ (Дата обращения 23.01.2025).
[14]. Веб-сайт https://github.com/brainteaser-ov/textgrid (Дата обращения 23.01.2025).
[15]. Graves, A., Mohamed, A., Hinton, G. Speech recognition with deep recurrent neural networks // International Conference on Acoustics, Speech and Signal Processing. 2013. pp. 6645-6649. Доступно по ссылке: https://arxiv.org/abs/1303.5778 (Дата обращения 23.01.2025).
[16]. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. r., Jaitly, N., Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups // IEEE Signal Processing Magazine, 29(6). 2012. pp. 82-97. Доступно по ссылке: https://www.cs.toronto.edu/~hinton/absps/DNN-2012-proof.pdf (Дата обращения 23.01.2025).
[17]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. Attention is all you need // Advances in Neural Information Processing Systems. 2017. pp. 5998-6008. Доступно по ссылке: https://arxiv.org/abs/1706.03762 (Дата обращения 23.01.2025).
[18]. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. pp. 4171–4186. Доступно по ссылке: https://aclanthology.org/N19-1423.pdf (Дата обращения 23.01.2025)
[19]. Cui, Y., Jia, M., Lin, T. Y., Song, Y., Belongie, S. Class-balanced loss based on effective number of samples // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 9268-9277. Доступно по ссылке: https://arxiv.org/abs/1901.05555 (Дата обращения 23.01.2025)
[20]. Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., Le, Q. V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition // Proc. Interspeech 2019. 2019. pp. 2613-2617. Доступно по ссылке: https://arxiv.org/abs/1904.08779 (Дата обращения 23.01.2025).
[21]. Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., Vinyals, O. Learning the speech front-end with raw waveform CLDNNs // Proc. Interspeech 2015. 2015. pp. 1-5. Доступно по ссылке: https://www.ee.columbia.edu/~ronw/pubs/interspeech2015-waveform_cldnn.pdf (Дата обращения 23.01.2025)

[22]. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks, Volume 18, Issues 5–6. 2005. pp. 602-610. doi.org/10.1016/j.neunet.2005.06.042 (Дата обращения 23.01.2025).
[23]. Bahdanau, D., Cho, K., Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate // Proc. ICLR 2015. 2015. Доступно по ссылке: https://arxiv.org/abs/1409.0473. (Дата обращения 23.01.2025)
[24]. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. // Journal of Artificial Intelligence Research. 16. 2002. pp. 321–357. Доступно по ссылке: http://dx.doi.org/10.1613/jair.953 (Дата обращения 23.01.2025).
[25]. Toshniwal, S., Bahdanau, D., Sagayama, S., Bengio, Y. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition // Proc. Interspeech 2017. 2017. pp. 3532-3536. Доступно по ссылке: https://arxiv.org/pdf/1704.01631(Дата обращения 23.01.2025).
[26]. Ioffe, S., Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // Proc. ICML 2015. 2015. pp. 448-456. Доступно по ссылке: https://arxiv.org/abs/1502.03167 (Дата обращения 23.01.2025).
[27]. Powers, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation // Journal of Machine Learning Technologies. 2(1). 2011. pp. 37–63. Доступно по ссылке: https://arxiv.org/abs/2010.16061 (Дата обращения 23.01.2025).
[28]. He, H., Garcia, E. A. Learning from Imbalanced Data // IEEE Transactions on Knowledge and Data Engineering. 21 (9). 2009. pp. 1263–1284. doi: 10.1109/TKDE.2008.239 (Дата обращения 23.01.2025).

Информация об авторах / Information about authors

Оксана Владимировна ГОНЧАРОВА – кандидат филологических наук, доцент, руководитель научно-образовательного центра «Интеллектуальный анализ данных» ФГБОУ ВО Пятигорский государственный университет, доцент кафедры русского языка и методики его преподавания ФГАОУ ВО Российский университет дружбы народов имени Патриса Лумумбы, старший научный сотрудник лаборатории Лингвистических платформ НИИ «Институт системного программирования им. В. П. Иванникова РАН» (техническая поддержка научно-исследовательской работы) с 2024 года. Сфера научных интересов: акустическая фонетика, просодия, социолингвистика, обработка естественного языка.

Oksana Vladimirovna GONCHAROVA – Cand. Sci. (Philology), Associate Professor, Head of the Scientific and Educational Center "Intellectual Data Analysis" of the Pyatigorsk State University, Associate Professor of the Department of Russian Language and Teaching Methods at Patrice Lumumba Peoples' Friendship University of Russia, Senior Researcher at the Laboratory of Linguistic Platforms of the Ivannikov Institute for System Programming of the Russian Academy of Sciences (technical support for research work) since 2024. Research interests: acoustic phonetics, prosody, sociolinguistics, natural language processing.