МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математического обеспечения вычислительных комплексов и
информационных систем

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ФОНЕМ НА ОСНОВЕ СОВРЕМЕННЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 — Математическое обеспечение и администрирование
информационных систем
профиль «Технологии программирования»
факультета компьютерных наук и информационных технологий
Родина Ивана Сергеевича

Научный руководитель
Д-р физ-мат. наук _____ Андрейченко Д. К.

Заведующий кафедрой
МОВКиИС _____ Андрейченко Д. К.

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы обусловлена быстрым развитием технологий автоматического распознавания речи и их широким применением — от голосовых помощников до систем обучения произношению. Одной из ключевых задач является точное распознавание фонем — минимальных звуковых единиц речи. Это важно для создания интерпретируемых и точных систем ASR (Automatic Speech Recognition), а также для задач лингвистического анализа и обучения произношению иностранных языков.

Цель данной работы — разработать систему, способную распознавать фонемы английского языка из аудиозаписей. Для этого планируется использовать современные методы обработки сигналов и машинного обучения, в том числе рекуррентные нейронные сети (RNN) и трансформеры (Transformer).

В рамках работы были поставлены следующие задачи:

1. Исследовать принципы обработки речевых сигналов.
2. Разобрать методы перевода речевых сигналов в текст.
3. Провести анализ существующих решений, которые используют методы глубокого обучения.
4. Разобрать фонетические алфавиты, которые используются для обозначения фонем.
5. Рассмотреть применение рекуррентных нейронные сетей и трансформеров для перевода речевых сигналов в фонемы
6. Провести сравнительный анализ моделей машинного обучения.
7. Найти и подготовить данные для моделей машинного обучения.
8. Спроектировать архитектуру веб-сервиса для работы пользователя с моделью.
9. Реализовать прототип веб-сервиса для демонстрации работы модели.
10. Протестировать систему на реальных аудиозаписях, выявить ограничения.

Методологические основы.

Проектирование и реализация системы автоматического распознавания фонем в данной работе основаны на современных подходах машинного обучения, цифровой обработки сигналов и лингвистического анализа. Методологическую основу составили труды российских исследователей: Алёшина Н. А., Акуратера Д., Барского А. Б., Гульчеева В. А., Карпова А. А., Кукушкина О. И., Мещерякова Р. В., Мясниковой Е. Н., Пойнтера Я., Поповой А., Потапова А.,

Рахманенко И. А., Созыкина А. В., Столбова М. Б., Тампеля И. Б., а также материалы МНМЦ УрФУ, освещающие вопросы обработки речи, архитектуры нейросетей и прикладной фонетики. Значительный вклад в методологическую основу внесли работы зарубежных исследователей: Алексея Бевски (Alexei Baevski) по самообучающимся моделям распознавания речи, Алекса Грейвза (Alex Graves) по рекуррентным нейронным сетям и CTC, Сеппа Хохрейтера (Sepp Hochreiter) по архитектуре LSTM, Ашиша Вашвани (Ashish Vaswani) по трансформерам, а также исследования Йошуа Бенджио (Yoshua Bengio) по обучению долгосрочных зависимостей в нейронных сетях.

Практическая значимость работы.

Практическая значимость заключается в возможности применения разработанной системы для анализа фонетического уровня речи, что может быть полезно в: образовательных платформах для изучающих иностранные языки, лингвистических исследованиях, системах речевого синтеза и ASR, логопедических приложениях.

Структура и объем работы.

Бакалаврская работа состоит из введения, 9 глав, заключения, списка использованных источников и 11 приложений. Основное содержание включает 69 страниц, 18 рисунков и подробное описание реализации модели и веб-сервиса. Объём списка литературы — 55 источников.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические основы обработки речевых сигналов» содержит изложение базовых понятий цифровой обработки речи: дискретизация, оконные функции, лог-мел-спектрограмма, мел-кепстральные коэффициенты. Описаны основные подходы к получению информативных признаков из аудиосигнала.

Второй раздел «Теоретические основы фонетической транскрипции» посвящён фонетике и фонологии. Приведены определения фонем, формант, а также сравнительный анализ систем транскрипции: IPA и ARPABET.

В разделе рассмотрены две основные системы фонетической транскрипции:

IPA (International Phonetic Alphabet) — международный фонетический алфавит, разработанный для универсальной записи звуков всех языков мира. Использует специальные символы, основанные на латинском и греческом алфавитах, с дополнительными диакритическими знаками.

ARPABET — фонетический алфавит, разработанный для английского языка в рамках проекта ARPA (Advanced Research Projects Agency). Использует ASCII-символы для представления фонем, что делает его удобным для компьютерной обработки.

Третий раздел «Фонетика и фонологическая структура речи» содержит описание фонологической структуры английского языка, акцентируя внимание на фонемах, важнейших для систем распознавания речи. Подробно рассмотрен словарь CMUdict.

В разделе даны определения ключевых понятий фонетики:

Фонема — это минимальная звуковая единица языка, которая может различать значения слов. Фонемы являются абстрактными единицами, представляющими классы звуков, которые носители языка воспринимают как одинаковые. Например, в английском языке звуки [p] и [b] являются разными фонемами, так как их замена может изменить значение слова (например, "pat"и "bat").

Форманты — это резонансные частоты голосового тракта, которые определяют тембр гласных звуков. В спектре речи форманты проявляются как пики амплитуды на определённых частотах. Первые три форманты (F1, F2, F3) наиболее важны для различения гласных звуков:

— F1 (первая форманта) — связана с высотой подъёма языка

— F2 (вторая форманта) — отражает продвинутость языка вперёд/назад
— F3 (третья форманта) — связана с формой губ и дополнительными резонансами

Четвёртый раздел «Методы обработки речи с помощью моделей глубокого обучения» представляет обзор нейросетевых архитектур: рекуррентных сетей (RNN, LSTM), трансформеров, а также механизма внимания и метода обучения CTC Loss, применяемого для расшифровки фонемных последовательностей.

Пятый раздел «Анализ предметной области» включает обзор существующих датасетов для распознавания речи, описание систем G2P, анализ моделей на основе RNN и Transformer, а также краткое сравнение подходов.

В ходе анализа были рассмотрены следующие системы G2P:

— CMUdict — словарно-правиловая система для английского языка
— Phonetisaurus — статистическая система на основе WFST
— Sequitur G2P — система на основе деревьев CART
— eSpeak — многоязычная система с ручными правилами
— G2P-seq2seq — нейросетевая система на основе Seq2Seq
— ESPnet G2P — система на основе трансформеров
— Festival — система на основе деревьев решений
— Open G2P — нейросетевая система на трансформерах
— G2P-модели fairseq — системы на основе трансформеров

Также были проанализированы следующие архитектуры нейронных сетей:

Рекуррентные сети (RNN):

— Стандартные RNN — базовые рекуррентные сети для обработки последовательностей
— LSTM — сети с длительной кратковременной памятью и тремя типами ворот
— GRU — облегченная версия LSTM с двумя воротами
— BiLSTM — двунаправленные LSTM для учета контекста с обеих сторон
— RNN-T — архитектура для потокового распознавания речи

Трансформеры:

— Wav2Vec2/HuBERT/WavLM — предобученные модели с CTC
— Мультиязычные трансформеры — модели для работы с несколькими языками

— Wav2Vec-U — самообучающиеся модели без размеченных данных
— Conformer — гибридные модели, сочетающие свертки и самовнимание

Шестой раздел «Подготовка данных» описывает процесс сбора и предобработки аудиоданных: преобразование текста в фонемные последовательности, извлечение признаков (мел-спектрограмм), реализация и тестирование VAD.

Для обучения моделей был использован корпус Mozilla Common Voice, включающий следующие версии:

— Common Voice Corpus 1
— Common Voice Corpus 2
— Common Voice Corpus 3
— Common Voice Corpus 4
— Common Voice Corpus 18.0
— Common Voice Corpus 19.0

Общий объем собранных данных составил 489 ГБ. В процессе предобработки из датасета были отобраны только необходимые столбцы: текст и аудиозаписи. Далее текст был преобразован в фонемную транскрипцию с использованием словаря CMUdict, а аудиозаписи — в лог-мел-спектрограммы для последующего обучения моделей.

Седьмой раздел «Реализация и сравнение моделей машинного обучения» содержит описание архитектуры моделей RNN и Transformer, использующих CTC Loss. Выполнена их реализация, обучение и сравнительный анализ качества распознавания фонем.

В работе были реализованы две архитектуры моделей сетей для распознавания фонем:

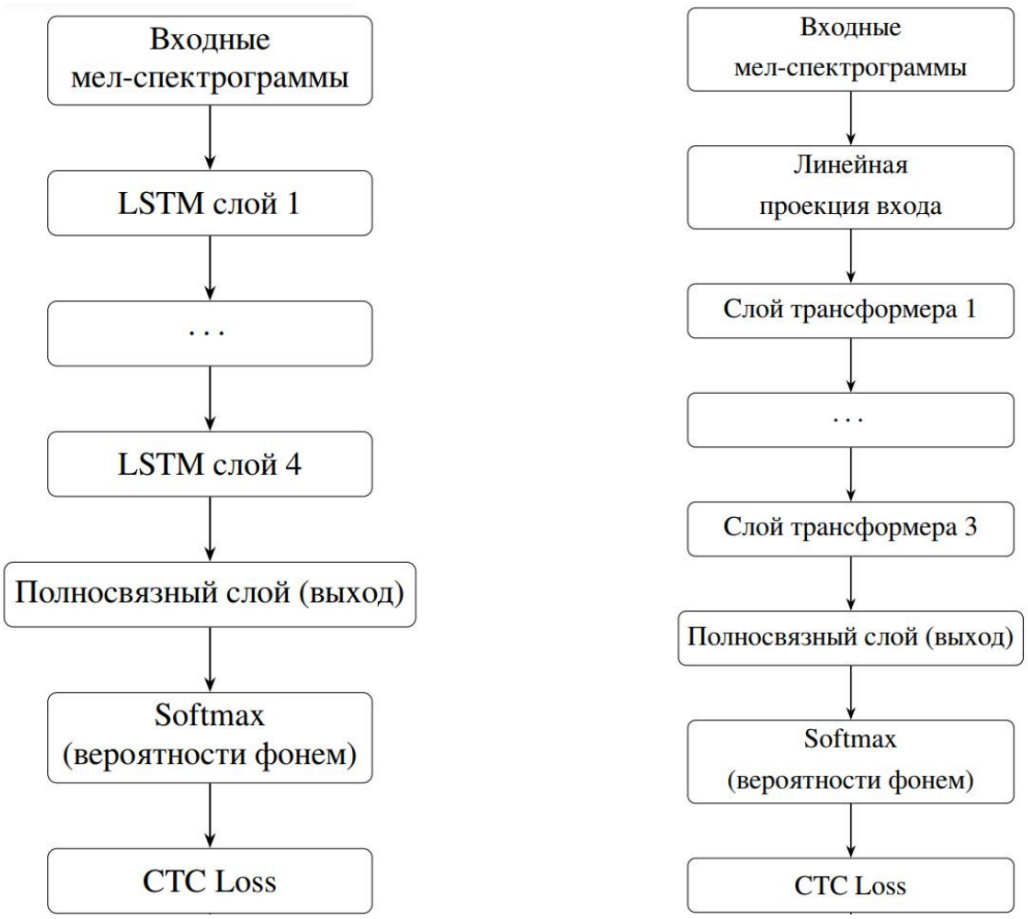

Рисунок 1 – Архитектуры моделей RNN и Transformer для распознавания фонем

Как видно на рисунке 1, обе модели используют CTC Loss для обучения и имеют схожую структуру входных и выходных данных, но различаются в способе обработки последовательностей.

В ходе экспериментального исследования были получены метрики качества обучения для обеих моделей, а также проведен анализ типичных ошибок распознавания. Ниже представлены результаты обучения по эпохам и примеры ошибок для обеих моделей.

Таблица 1 – Обучение RNN

Эпоха	CTC Loss	PER, %
1	30.0	90.0
20	10.5	82.4
100	5.8	75.1
200	4.2	68.3
300	1.4	62.0
400	1.3	56.2
500	1.3	51.0
600	1.2	46.5
700	1.2	42.8
1000	1.1	40.1

Таблица 2 – Обучение Transformer

Эпоха	CTC Loss	PER, %
1	35.0	92.0
10	15.5	85.4
20	8.8	78.1
30	5.2	72.3
50	3.4	68.0
100	2.6	62.2
200	2.4	58.0
300	2.1	54.5
400	1.6	51.8
500	1.5	48.1

Восьмой раздел «Модификация токенизатора Whisper» включает описание изменений, внесённых в токенизатор модели Whisper для поддержки фонемной транскрипции, а также этапов обучения и оценки качества.

Whisper представляет собой многоязычную модель распознавания речи, основанную на архитектуре трансформера. В данной работе была предпринята попытка модифицировать токенизатор для работы с фонемами ARPABET, что позволило бы использовать предобученную модель для фонемной транскрипции без существенных изменений в архитектуре.

Модификация токенизатора включала следующие изменения:

1. Добавление всех фонем ARPABET в словарь токенизатора
2. Модификация правил токенизации для корректной обработки фонемных последовательностей
3. Настройка специальных токенов для работы с фонемами

Таблица 3 – Метрики качества по эпохам обучения

Эпоха	PER	CTC Loss
1	28.5%	0.423
2	28.2%	0.421
3	28.7%	0.425
4	28.3%	0.422
5	28.4%	0.424
6	28.6%	0.423
7	28.3%	0.421
8	28.5%	0.422
9	28.4%	0.423
10	28.5%	0.424

Как видно из таблицы 3, метрики качества демонстрируют стабильное поведение с небольшими колебаниями. PER колеблется в районе 28.5%, а CTC Loss - около 0.423. Такое поведение метрик указывает на то, что модель достигла определённого уровня производительности, и дальнейшее обучение не приводит к существенному улучшению.

Анализ результатов показал следующие преимущества модификации токенизатора:

1. Сохранение предобученных весов модели
2. Улучшение качества фонемной транскрипции
3. Стабильность в процессе обучения
4. Возможность использования существующей архитектуры

Девятый раздел «Разработка веб-интерфейса» описывает клиент-серверное взаимодействие, реализацию интерфейса на Flutter, интеграцию модели и VAD в WebAssembly-модуле и общую архитектуру сервиса.

Веб-интерфейс был реализован с использованием Flutter для обеспечения кросс-платформенности и современного пользовательского опыта. Основной функционал включает загрузку аудиофайлов, их обработку с помощью VAD для выделения речевых сегментов и последующее распознавание фонем с использованием обученных моделей. Интерфейс предоставляет пользователю возможность выбора модели (RNN или Transformer) для распознавания, а также отображает результаты в удобном для анализа формате.

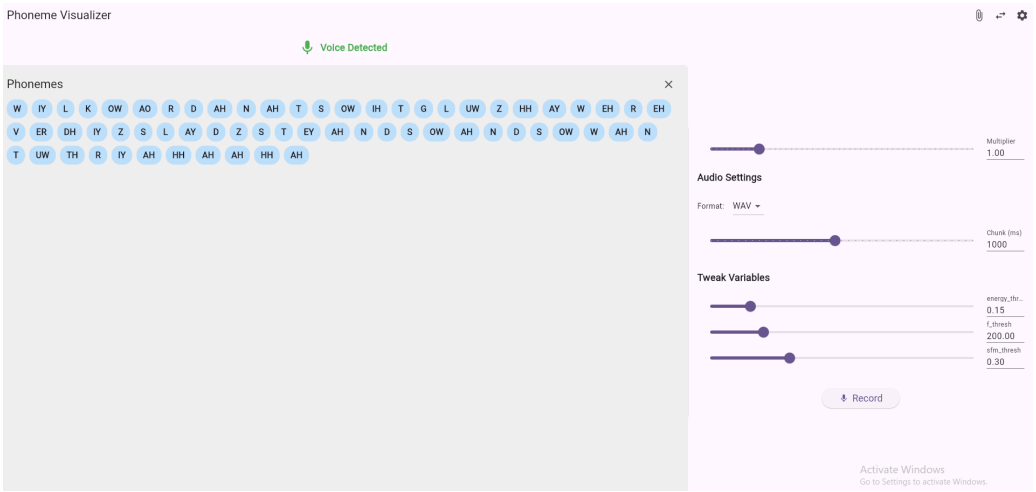

Рисунок 2 – Веб-интерфейс системы распознавания фонем

Как видно на рисунке 2, интерфейс разделен на несколько функциональных зон: область загрузки аудиофайла, панель управления с выбором модели и параметров обработки, а также область отображения результатов распознавания. Результаты представляются в виде последовательности фонем с возможностью сравнения с эталонной транскрипцией, что позволяет пользователю оценить качество работы системы.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы была поставлена и решена задача распознавания фонем английской речи с использованием методов глубокого обучения. Проведён комплексный анализ существующих подходов к обработке речевых сигналов, рассмотрены фонетические алфавиты IPA и ARPABET, изучены архитектуры современных нейронных сетей и проведена их оценка применительно к задаче фонемного распознавания.

В рамках работы была проведена тщательная подготовка данных на основе датасета Mozilla Common Voice. Данные были отфильтрованы по региону происхождения (England) и отобраны записи с положительной оценкой качества. Для обеспечения достаточного объема данных были использованы шесть версий корпуса (Common Voice 1.0-4.0, 18.0, 19.0). Реализованный алгоритм VAD показал точность до 94.5% и полноту 92.0%.

Особое внимание было уделено преобразованию текста в фонемную форму с использованием CMUdict. Разработанный пайплайн включал приведение текста к нижнему регистру, очистку от знаков препинания, токенизацию и поиск в словаре, а также обработку множественных вариантов произношения.

Были реализованы и протестированы три различных подхода к распознаванию фонем. Рекуррентные нейронные сети (RNN) с архитектурой LSTM достигли PER 40.1% после 1000 эпох обучения, используя пятислойную архитектуру с 256 нейронами в каждом направлении. Трансформеры показали результат PER 48.1% после 500 эпох обучения, с архитектурой из 4 слоев энкодера и использованием многоголового внимания. Адаптация Whisper показала стабильный PER порядка 28.5%, что является наилучшим результатом среди всех подходов.

Разработан полноценный веб-сервис с современной архитектурой: фронтенд на Flutter с кроссплатформенной поддержкой, бэкенд на Python FastAPI, реализация VAD на WebAssembly (Rust). Пользовательский интерфейс разработан с учетом современных требований к UX и включает визуализацию результатов распознавания.

Проведённая работа позволила достичь значительных результатов. Была создана эффективная система предобработки аудио с точностью VAD до 94.5%, реализованы и протестированы три различных подхода к распознаванию фонем. Наилучший результат PER 28.5% был достигнут с использованием адап-

тированной модели Whisper, что демонстрирует эффективность предобученных моделей при правильной адаптации. Рекуррентные сети и трансформеры показали схожие результаты, что говорит о том, что обе архитектуры могут быть эффективно использованы для задачи фонемного распознавания, хотя и требуют дополнительной оптимизации для достижения лучших результатов.

В качестве направлений дальнейшего развития системы можно выделить улучшение архитектуры трансформера с учетом специфики фонемного распознавания, оптимизацию рекуррентных сетей для использования на мобильных устройствах, исследование гибридных подходов, комбинирующих преимущества различных архитектур. Также планируется расширение языковой поддержки системы и улучшение качества распознавания на неидеальных аудиозаписях.

Результаты работы демонстрируют, что наилучшие результаты достигаются при адаптации предобученных моделей, таких как Whisper, хотя это требует тщательной настройки и модификации архитектуры. Рекуррентные сети и трансформеры показали схожие результаты, что говорит о том, что обе архитектуры могут быть эффективно использованы для задачи фонемного распознавания, хотя и требуют дополнительной оптимизации для достижения лучших результатов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Mel frequency cepstral coefficient: a review / Shalbbya Ali, Safdar Tanweer, Syed Sibtain Khalid, Naseem Rao // *ICIDSSD*. — 2020.

2 *Алёшин, Никита Александрович*. Рекуррентные нейронные сети / Никита Александрович Алёшин // World science: problems and innovations. — 2021. — С. 10–12.

3 *Soydaner, Derya*. Attention Mechanism in Neural Networks: Where it Comes and Where it Goes / Derya Soydaner. — 2022.

4 *Microsoft*. Microsoft Azure Cognitive Services Documentation: Speech-to-Text Service. — https://learn.microsoft.com/azure/cognitive-services/speech-service/. — 2023.

5 wav2vec-U: Self-Supervised Unsupervised Speech Recognition / Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, Michael Auli // *arXiv preprint arXiv:2105.11084*. — 2022. https://arxiv.org/abs/2105.11084.

6 *Bengio, Yoshua*. Learning long-term dependencies with gradient descent is difficult / Yoshua Bengio, Patrice Simard, Paolo Frasconi // *IEEE transactions on neural networks*. — 1994. — Т. 5, № 2. — С. 157–166.

7 *Bernard, Mathieu*. Phonemizer: Text to Phones Transcription for Multiple Languages in Python / Mathieu Bernard, Hadrien Titeux // *Journal of Open Source Software*. — 2021. — Т. 6, № 68. — С. 3958. https://doi.org/10.21105/joss.03958.

8 *Goodfellow, Ian*. Deep Learning / Ian Goodfellow, Yoshua Bengio, Aaron Courville. — MIT Press, 2016.

9 Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation / Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre и др. // *arXiv preprint arXiv:1406.1078*. — 2014.

10 *University, Carnegie Mellon*. The CMU Pronouncing Dictionary. — 2020. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

11 *Duddington, Jonathan*. eSpeak: Speech Synthesizer. — 2020. http://espeak.sourceforge.net/.

12 ESPnet: End-to-End Speech Processing Toolkit / Shinji Watanabe, Takaaki Hori, Shigeki Karita и др. // Interspeech. — 2018. — С. 2207–2211.

13 fairseq: A Fast, Extensible Toolkit for Sequence Modeling / Myle Ott, Sergey Edunov, Alexei Baevski и др. // Proceedings of NAACL-HLT 2019: Demonstrations. — 2019. — С. 48–53.

14 *Black, Alan W.* The Festival Speech Synthesis System. — 1999. http://www.cstr.ed.ac.uk/projects/festival/.

15 *roedoejet*. G2P: Grapheme to Phoneme Conversion. — https://github.com/roedoejet/G2P. — 2023. — GitHub repository for grapheme to phoneme conversion.

16 Grapheme-to-Phoneme Conversion with Sequence-to-Sequence Models / Kanishka Rao, Fuchun Peng, Haşim Sak, Françoise Beaufays // Interspeech. — 2015. — С. 3330–3334.

17 *Schultz, Tanja*. Language-independent and language-adaptive acoustic modeling for speech recognition. — 2001.

18 Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks / Alex Graves, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber // *Proceedings of the 23rd International Conference on Machine Learning*. — 2006. — С. 369–376.

19 *Graves, Alex*. Sequence Transduction with Recurrent Neural Networks / Alex Graves // *arXiv preprint arXiv:1211.3711*. — 2012.

20 *Graves, Alex*. Speech Recognition with Deep Recurrent Neural Networks / Alex Graves, Abdel-rahman Mohamed, Geoffrey Hinton // ICASSP. — 2013. — Pp. 6645–6649.

21 Conformer: Convolution-augmented Transformer for Speech Recognition / Anmol Gulati, James Qin, Chung-Cheng Chiu и др. // *arXiv preprint arXiv:2005.08100*. — 2020.

22 *Hochreiter, Sepp*. Long Short-term Memory / Sepp Hochreiter, Jürgen Schmidhuber // *Neural Computation*. — 1997. — Т. 9, № 8. — С. 1735–1780.

23 *Karpathy, Andrej*. Deep visual-semantic alignments for generating image descriptions / Andrej Karpathy, Li Fei-Fei // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — С. 3128–3137.

24 Streaming Conformer-based Speech Recognition Using Non-causal Convolution and Causal Self-attention / Jaeyoung Kim, Minjae Kim, Dongsuk Park и др. // *arXiv preprint arXiv:2110.09705*. — 2022.

25 *Panayotov, Vassil*. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. — arXiv:1504.08755 [cs.CL]. — 2015. https://arxiv.org/abs/1504.08755.

26 Wav2vec-U 2.0: Learning Unsupservised Speech Recognition from Multilingual Audio / Yatharth Saraf Liu, Alexei Baevski, Alexis Conneau, Michael Auli // *arXiv preprint arXiv:2204.02492*. — 2022. https://arxiv.org/abs/2204.02492.

27 Understanding LSTM networks[Электронный ресурс]. — URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Загл. с экр. Яз. англ.

28 LSTM: A Search Space Odyssey / Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník и др. // *IEEE Transactions on Neural Networks and Learning Systems*. — 2015. — Т. 28. — С. 2222–2232.

29 *Hochreiter, Sepp*. Long Short-term Memory / Sepp Hochreiter, Jürgen Schmidhuber // *Neural computation*. — 1997. — Т. 9. — С. 1735–80.

30 *Medin, Lou B.* Self-Supervised Models for Phoneme Recognition: Applications in Children's Speech for Reading Learning / Lou B. Medin, Thomas Pellegrini, Laure Gelin // Proceedings of Interspeech 2024. — 2024. — С. 4181–4185.

31 *ASR-project*. Multilingual Phoneme Recognition. — https://github.com/ASR-project/Multilingual-PR. — 2023. — GitHub repository.

32 *Мясникова, Е.Н.* Объективное распознавание звуков речи / Е.Н. Мясникова. — Tbilisi State University, 2013. https://books.google.ru/books?id=M7X-AgAAQBAJ.

33 *AI, Coqui*. Open G2P. — 2022. https://github.com/coqui-ai/TTS.

34 *Novak, Josef R.* Phonetisaurus: A WFST-driven Phoneticizer - Framework Improvements and Scale-Up / Josef R. Novak, Nobuaki Minematsu, Keikichi Hirose // Proceedings of the 2016 Conference of the North American

Chapter of the Association for Computational Linguistics: Demonstrations. — 2016. — С. 1–5.

35 *Пойнтер, Я.* Программируем с PyTorch: Создание приложений глубокого обучения / Я. Пойнтер, Д. Акуратер, А. Попова. — Питер, 2024. https://books.google.ru/books?id=xzomEQAAQBAJ.

36 *Рахманенко, Иван Андреевич.* Анализ идентификационных признаков в речевых данных с помощью GMM-UBM системы верификации диктора / Иван Андреевич Рахманенко, Роман Валерьевич Мещеряков // *Информатика и автоматизация*. — 2017. — Vol. 3, no. 52. — Pp. 32–50.

37 Efficient Voice Activity Detection Algorithms Using Long-term Speech Information / Javier Ramirez, Jose C. Segura, Carmen Benitez и др. // Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009). — Glasgow, Scotland: EURASIP, 2009. — С. 1345–1349. — A simple and efficient energy-based VAD algorithm that uses adaptive thresholds and temporal smoothing. https://eurasip.org/Proceedings/Eusipco/Eusipco2009/contents/papers/1569192958.pdf.

38 *Rao, Kanishka.* Exploring Architectures, Data and Units For Streaming End-to-End Speech Recognition with RNN-Transducer / Kanishka Rao, Hasim Sak, Rohit Prabhavalkar // *2018 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. — 2018. — С. 93–98.

39 *DiPietro, Robert.* Chapter 21 – Deep learning: rNNs and LSTM / Robert DiPietro, Gregory D. Hager // Handbook of Medical Image Computing and Computer Assisted Intervention / Под ред. S. Kevin Zhou, Daniel Rueckert, Gabor Fichtinger. — Academic Press, 2020. — The Elsevier and MICCAI Society Book Series. — С. 503–519.

40 *Schuster, Mike.* Bidirectional Recurrent Neural Networks / Mike Schuster, Kuldip K. Paliwal // *IEEE Transactions on Signal Processing*. — 1997. — Т. 45, № 11. — С. 2673–2681.

41 *Bisani, Maximilian.* A joint sequence model for grapheme-to-phoneme conversion / Maximilian Bisani, Hermann Ney // Interspeech. — 2008. — С. 117–120.

42 *Shim, Kwangyoun*. A Comparison of Transformer, Convolutional, and Recurrent Neural Networks on Phoneme Recognition / Kwangyoun Shim, Wonyong Sung // *arXiv preprint arXiv:2210.00367*. — 2022. https://arxiv.org/abs/2210.00367.

43 *Созыкин, Андрей Владимирович*. Обзор методов обучения глубоких нейронных сетей / Андрей Владимирович Созыкин // *Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика*. — 2017. — Vol. 6, no. 3. — Pp. 28–59.

44 *R, Srinath N*. IPA-Wav2Vec2-Phoneme-Recognition. — https://github.com/Srinath-N-R/IPA-Wav2Vec2-Phoneme-Recognition. — 2023.

45 *Столбов, Михаил Борисович*. Основы анализа и обработки речевых сигналов / Михаил Борисович Столбов. — федеральное государственное автономное образовательное учреждение высшего образования, 2021.

46 *Sutskever, Ilya*. Sequence to sequence learning with neural networks / Ilya Sutskever, Oriol Vinyals, Quoc V Le // Advances in neural information processing systems. — Т. 27. — 2014.

47 *Тампель, ИБ*. Автоматическое распознавание речи / ИБ Тампель, Алексей Анатольевич Карпов // *Учебное пособие.- СПб: Университет ИТМО*. — 2016.

48 *Rousseau, Anthony*. TED-LIUM: an Automatic Speech Recognition dedicated corpus. — Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). — 2012.

49 *Linguistic Data Consortium*. TIMIT Acoustic-Phonetic Continuous Speech Corpus. — https://catalog.ldc.upenn.edu/LDC93S1. — 1993. — LDC93S1.

50 *Кукушкин, О. И.* Трансформеры в обработке естественного языка: архитектура и применение / О. И. Кукушкин // *Вопросы кибернетики*. — 2021. — № 3. — С. 45–62.

51 *Vaswani, Ashish*. Attention is all you need / Ashish Vaswani и др. // Advances in neural information processing systems. — Т. 30. — 2017.

52 *OpenAI Community*. Transcribe to IPA (International Phonetic Alphabet) – Whisper GitHub Discussions. — https://github.com/openai/whisper/discussions/. — 2023.

53 *Linguistic Data Consortium*. CSR-I (WSJ0) Complete. — https://catalog.ldc.upenn.edu/LDC93S6A. — 1993. — LDC93S6A.

54 *Xu, Qiantong*. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition / Qiantong Xu, Alexei Baevski, Michael Auli // Proceedings of Interspeech 2022. — 2022. — C. 2111–2115.

55 *Yen, Hsien-Chi*. Boosting End-to-End Multilingual Phoneme Recognition through Exploiting Universal Speech Attributes Constraints / Hsien-Chi Yen, Sabato Marco Siniscalchi, Chin-Hui Lee // *arXiv preprint arXiv:2309.08828*. — 2023. https://arxiv.org/abs/2309.08828.