УДК 004.522

ПОСТРОЕНИЕ ВЫСОКОТОЧНОЙ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ТУРКМЕНСКОГО ЯЗЫКА НА ОСНОВЕ КОМБИНИРОВАННОГО ТРАНСФОРМЕРНОГО ПОДХОДА

М.Т. Мырадов
Заведующий кафедры «Информационные системы»
Институт Телекоммуникаций и информатики Туркменистана,
maksat.myradow.92@mail.ru

М.Т.Мырадов
Окончил Института телекоммуникаций и информатики Туркменистана. Область научных интересов распознавания речи, искусственный интеллект, защита данных

Аннотация: Современные системы распознавания речи на базе моделей Whisper Large-v3 и Llama-3 используют архитектуру Transformer и механизмы семантического прогнозирования для высокоточной интерпретации акустических сигналов в текстовую последовательность. Применение оптимизированных алгоритмов Faster-Whisper и архитектуры синтеза VITS позволяет эффективно обрабатывать специфические особенности туркменского языка, обеспечивая при этом высокую скорость вычислений и естественность звучания. Экспериментальное обучение модели на массиве из 4000 часов аудиоданных с использованием графических процессоров NVIDIA A100 подтвердило значительное снижение потерь и достижение высокого уровня точности распознавания.

Ключевые слова: Распознавание речи, языковые модели, архитектура Transformer, Whisper Large-v3, Llama-3, глубокое машинное обучение, семантический анализ, синтез речи, туркменский язык.

Введение: Современная научная интерпретация использования языковых моделей в распознавании речи основывается на концепциях семантического прогнозирования и мультимодального синтеза.

В настоящее время модели в системах STT (Speech-to-Text) эволюционировали в интеллектуальные процессоры, которые не просто идентифицируют слова, но и глубоко анализируют контекст высказывания. Основная задача таких моделей, как Whisper Large-v3 и Llama-3, заключается в интерпретации акустических сигналов посредством вероятностного анализа последовательностей слов. Благодаря архитектуре Transformer, нейронная сеть способна аппроксимировать логические характеристики предложения даже в условиях акустических помех или высокого уровня шума. Интеграция языковых моделей непосредственно в процесс декодирования звука позволяет в реальном времени корректировать грамматические и семантические ошибки.

Применение Whisper Large-v3 и Llama-3 имеет критическое значение для таких языков, как туркменский. Современное распознавание речи, сочетающее спектральный анализ и глубокое машинное обучение, обеспечивает высокую точность интерпретации данных [1].

Модель Whisper Large-v3 разработана для автоматического распознавания речи (ASR) и преобразования голоса в текст. Llama-3-8B, в свою очередь, является генеративной языковой моделью, работающей исключительно с текстовыми данными. Обе модели базируются на архитектуре Transformer, однако применяются к различным модальностям данных. Whisper

Large-v3 обрабатывает аудиосигналы, используя архитектуру Encoder-Decoder, тогда как Llama-3-8B представляет собой Decoder-only модель для работы с текстом [2]. Процесс обработки начинается с этапа сегментации речи, где входной сигнал преобразуется в логарифмическую мел-спектрограмму. Использование 128 спектральных фильтров в версии Large-v3 обеспечивает высокоточную передачу звуковых характеристик и частотных параметров речи. Энкодер состоит из двух высокоточных слоев с использованием сверточных фильтров и функции активации GELU (Gaussian Error Linear Unit). Эти слои сокращают временную размерность аудиосигнала вдвое. Затем данные передаются в 32 блока Трансформера. Каждый блок задействует механизм многоголового внимания (Multi-Head Attention), вычисляющий зависимости между всеми сегментами речевого тракта [3].

Декодер функционирует в авторегрессионном режиме, последовательно предсказывая текстовые токены. Он использует механизм перекрестного внимания (Cross-Attention) для фокусировки на релевантных признаках из энкодера на основе уже сгенерированного текста. Общее количество параметров модели составляет приблизительно 1,55 миллиарда [2].

Одной из ключевых особенностей является использование вращательных позиционных вложений (Rotary Positional Embeddings, RoPE). С математической точки зрения это реализуется путем вращения векторов запросов (query) и ключей (key) в комплексном пространстве с использованием тригонометрических функций. Данный механизм позволяет модели учитывать относительные расстояния между словами в больших контекстных окнах без потери качества обработки.Для нормализации данных на каждом слое применяется RMSNorm (Root Mean Square Layer Normalization).

В отличие от стандартной нормализации, данный метод не вычитает среднее значение, а ограничивается делением вектора на его среднеквадратичную норму. Это существенно ускоряет вычислительные процессы и оптимизирует эффективность распознавания речи [4]. В данной архитектуре используется механизм Grouped-Query Attention (GQA). Этот подход значительно сокращает объем памяти, необходимый для хранения кэша (KV-cache), и ускоряет генерацию ответов.

Для активации слоев применяется функция SwiGLU, которая объединяет свойства функций Swish и GLU (Gated Linear Unit), обеспечивая более стабильное и эффективное обучение нейронной сети. Основное различие между моделями заключается в математической формулировке задачи и методах обработки входных данных (таблица 1).

Таблица 1. Основные различия между моделями

Разделы	**Whisper Large-v3**	**Llama-3-8B**	**Тип данных**	**Тренировочные данные**
Основная задача	Распознавание речи	Генерация текста	Аудио – Размеченный текст	Supervised
Архитектура	Трансформер типа «Только Декодер»	Декодер Трансформер	Спектрограммы Токены	Прогнозирование следующего токена

В частности, Whisper Large-v3 на начальном этапе выполняет кратковременное преобразование Фурье (STFT) для перевода аудиосигнала x(t) во временно-частотное представление.

В результате полученная спектрограмма X(\omega, t) передается в энкодер **Whisper Large-v3**, который вычисляет скрытые представления, используя механизм самовнимания (**Self-Attention**):

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/d_k)V$$

Декодер Whisper применяет маскированное самовнимание для генерации текста и механизм перекрестного внимания для синхронизации текстовых токенов с акустическими признаками, полученными из энкодера.

Напротив, модель **Llama-3-8B** решает задачу вероятностного моделирования текстовых последовательностей. Рассмотрим последовательность токенов x_1, x_2, \dots, x_t. Модель прогнозирует вероятность появления следующего токена:

$$p(x_{t+1} \; x_{<1}) = \text{softmax}\}(H^{L} \; W_{o})$$

Таким образом, Whisper оптимизирован для задачи выравнивания (alignment) звука и текста, тогда как Llama-3-8B ориентирована на авторегрессионную генерацию. Модель принимает на вход последовательность текстовых токенов и обучается предсказывать последующий элемент [5].

Подготовка и оптимизация системы STT MOTOR: Faster-Whisper представляет собой оптимизированную версию оригинальной модели OpenAI, которая функционирует значительно быстрее за счет квантования вычислений без существенной потери точности.

Данная модификация модели способна эффективно обрабатывать специфические данные, такие как медицинская терминология или редкие диалекты. Такая оптимизация позволяет системе более точно распознавать наречия, в которых стандартные модели часто допускают ошибки [6].

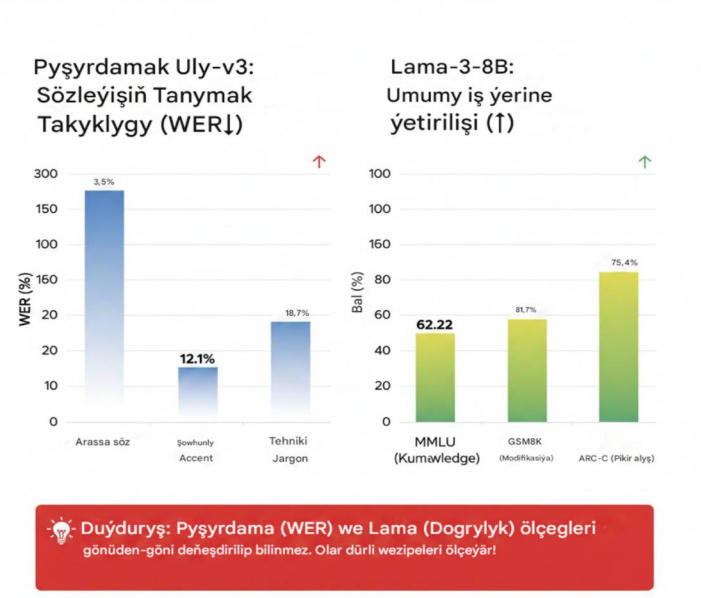

Рисунок 1. Whisper Large-v3 и Llama-3

Для оценки производительности данной модели мы используем метрику METEOR, которая обеспечивает более точный анализ по сравнению с простым сопоставлением слов.

Она распознает синонимы и учитывает корректность порядка слов, что делает оценку более объективной и приближенной к человеческому восприятию.

В отличие от стандартных методов, эта метрика признает, что замена слова синонимом не является грубой ошибкой.

Использование METEOR в сочетании с тонко настроенной моделью «Faster-Whisper» помогает разработчикам адаптировать систему для передачи глубокого смысла высказываний, а не просто набора звуков.

Логика Python: Логика программирования на языке Python представляет собой совокупность алгоритмических и структурных принципов, обеспечивающих корректную обработку данных и управление программными потоками.

Python использует высокоуровневые структуры, такие как условные операторы и функции, которые улучшают читаемость кода и формальную верификацию.

Благодаря динамической типизации и обширным стандартным библиотекам, язык эффективно применяется в научных вычислениях и системах искусственного интеллекта.

Логическая модель Python способствует быстрому прототипированию и минимизирует вероятность логических ошибок, что делает его оптимальным инструментом для построения сложных программных комплексов.

TTS MOTOR: VITS (End-to-End): VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) – это современная архитектура синтеза речи. Модель объединяет вариационный автокодировщик и генеративно-состязательное обучение для генерации речевого сигнала.

В отличие от каскадных систем TTS, VITS обучается напрямую без промежуточных акустических представлений, что снижает накопление ошибок.

Это делает VITS перспективным решением для масштабируемых систем синтеза речи.

Speaker: Компонент Speaker в системах TTS (Text-to-Speech) отвечает за моделирование индивидуальных характеристик голоса диктора.

Он реализуется в виде эмбеддингов (embeddings), которые отображают тембр, интонацию и особенности артикуляции речи.

Использование идентификаторов диктора позволяет осуществлять многоголосый синтез в рамках одной модели.

В научных исследованиях данный подход применяется для клонирования голоса и персонализации речевых интерфейсов.

Правильное моделирование компонента Speaker значительно повышает естественность синтезированной речи.

Заключение. Для обучения и запуска новой туркменской языковой модели мы использовали высокопроизводительные графические процессоры, в частности GPU: 2x NVIDIA A100 (80 ГБ VRAM).

```
[15] def load_audio(audio_path):
0s       # Load the audio file
         waveform, sample_rate = torchaudio.load(audio_path)

         # If the sample rate of the audio file is different, resample it
         if sample_rate != 16000:  # Whisper models expect 16kHz audio
             waveform = torchaudio.transforms.Resample(orig_freq=sample_rate, new_freq=16000)(waveform)

         return waveform.squeeze(0), 16000  # Return waveform and target sample rate
```

Обучение новой модели Whisper проводилось на основе 4000 часов аудиозаписей на туркменском языке в течение 14 дней (показатель Training Loss был снижен до 0.04).

Список литературы

[1] Надежное распознавание речи с помощью крупномасштабного слабого контроля Алек Рэдфорд, Чон Вук Ким, Тао Сюй, Грег Брокман, Кристин Макливи, Илья Суцкевер 2019

[2] Эффективное многоголовочное декодирование для распознавания речи на основе трансформеров Яэль Сегал-Фельдман, Авив Шамсян, Авив Навон, Гилл Хетц, Джозеф Кешет 2022

[3] Соединение кодировщика речи и большой языковой модели для автоматического распознавания речи. Вэньи Ю,Чанли Тан,Гуанчжи Сунь, Сяньчжао Чен, Тянь Тан, Вэй Ли, ,Цзэцзюнь Ма,Чао Чжан 2020

[4] B.H. Juang, and S. Furui. Automatic Recogni-tion and Understanding of Spoken Language–A First Step Toward Natural Human Machine Communication. Proc. IEEE, 88, No. 8, 2016, pp. 1142-1165.

[5] Hiroaki Sakoe and Seibi Chiba. Dynamic Pro-gramming Algorithms Optimization for Spoken Word Recoginition. IEEE Trans on acoustics, speech and sig-nal processing, vol.ASSP-26, no.1, december 2017.

[6] L.R.Rabiner and B.H.Juang. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 2016.

Авторский вклад

Мырадов Максат Тачмухаммедович – авторский вклад в архитектуру Transformer для преобразования голоса в текст, разработав, например, новый метод адаптации механизма внимания, оптимизирующий процесс обучения или интегрировав инновационные слои для улучшения обработки аудиоданных и повышения точности распознавания речи.

DEVELOPMENT OF A HIGH-PRECISION INTELLIGENT NATURAL LANGUAGE PROCESSING SYSTEM FOR THE TURKMEN LANGUAGE BASED ON A COMBINED TRANSFORMER APPROACH

M.T. Myradov

Head of the Department of Information Systems, The Institute of Telecommunications and Informatics

Abstract: Modern speech recognition systems based on Whisper Large-v3 and Llama-3 models utilize the Transformer architecture and semantic prediction mechanisms for high-precision interpretation of acoustic signals into text sequences. The application of optimized Faster-Whisper algorithms and the VITS synthesis architecture enables efficient processing of the specific features of the Turkmen language, while ensuring high computational speed and natural sound quality. Experimental training of the model on a dataset of 4000 hours of audio data using NVIDIA A100 GPUs confirmed a significant reduction in loss and the achievement of a high level of recognition accuracy.

Keywords: Speech recognition, language models, Transformer architecture, Whisper Large-v3, Llama-3, deep machine learning, semantic analysis, speech synthesis, Turkmen language.