**УДК 004.8**

# РАЗРАБОТКА ИНТЕГРАЛЬНОЙ СИСТЕМЫ ДЛЯ ОБРАБОТКИ И ПОНИМАНИЯ ЕСТЕСТВЕННОЙ РЕЧИ

**Дутов Д.А.[1], Митрофанов А.А.[2]**
**Научный руководитель – Митрофанов А.А.[2]**
[1]Университет ИТМО, [2]ООО "ЦРТ-Инновации"
maunberg@bk.ru

## Введение

В современном этапе развития информационных технологий методы обработки естественного языка (NLP, Natural Language Processing) занимают ключевое место в построении интеллектуальных систем. Существенным направлением развития NLP является понимание естественной речи (SLU, Speech Language Understanding), объединяющее методы анализа акустического сигнала и интерпретации его семантического содержания. Интеграция речевых и языковых компонентов позволяет создавать системы, способные к полноценному взаимодействию с человеком на естественном языке.

Модели SLU находят применение в широком спектре практических задач: от разработки голосовых ассистентов и диалоговых систем до автоматического перевода, мультимодального поиска, анализа клиентских обращений и мониторинга пользовательского опыта. В условиях стремительного роста объёмов аудиоданных особую актуальность приобретает создание универсальных архитектур, способных эффективно обрабатывать аудио-текстовые запросы различной природы.

Целью данной работы является разработка универсальной модели для обработки аудио-текстовых запросов, обеспечивающей адаптивность к различным акустическим и языковым сценариям и демонстрирующей устойчивое качество при решении широкого круга задач понимания речи.

## Основная часть

На начальном этапе разработки модели был определён целевой бенчмарк – Dynamic SuperB Phase 2, представляющий собой комплексную оценочную среду, включающую 180 задач из различных доменов: автоматической обработки речи, анализа аудиосигналов и музыкальной информации. Бенчмарк организован в виде развёрнутой таксономии, где каждая обобщённая задача декомпозируется на набор более специализированных подзадач. Такая структура формирует древовидный граф, листьями которого являются конкретные тестовые выборки. Иерархическая организация задач обеспечивает систематическое покрытие широкого спектра акустических и языковых сценариев и позволяет формировать структурированную обучающую выборку, ориентированную на многоаспектную оценку модели.

Разработанная модель имеет модульную архитектуру и включает следующие ключевые компоненты:

1. Речевой энкодер WavLM [1] — глубокая нейросетевая модель извлечения акустических представлений, предварительно обученная на масштабных речевых корпусах и предназначенная для формирования информативных признаков аудиосигнала.
2. Проекционный слой — адаптационный модуль, осуществляющий линейное или нелинейное преобразование акустических представлений в пространство признаков, согласованное с входным форматом языковой модели.
3. Языковая модель Qwen 2.5 7B Instruct [2] — крупная предобученная

трансформерная модель, ориентированная на обработку естественного языка и инструкционное следование.

4. Адаптер LoRA (Low-Rank Adaptation) [3] — метод параметрически эффективного дообучения больших языковых моделей, позволяющий адаптировать модель за счёт добавления низкоранговых матриц без обновления полного набора весов.

В процессе обучения параметры речевого энкодера и языковой модели оставались замороженными; обновлялись исключительно веса проекционного слоя и адаптера LoRA. Такой подход существенно снижает вычислительные затраты и требования к объёму обучающих данных, одновременно сохраняя генеративные и обобщающие способности предобученной языковой модели.

Для оценки производительности использовались модели-судьи — специализированные нейросетевые системы, обученные на задаче автоматической оценки качества сгенерированных ответов. С целью обеспечения объективности была сформирована сбалансированная тестовая выборка, включающая более 2500 аудиозаписей, вручную аннотированных экспертами. Набор охватывал широкий спектр сценариев (речевых, паралингвистических и акустических), что позволило всесторонне оценить корректность, содержательность и релевантность ответов модели, а также выбрать наиболее надёжную модель-судью.

После проведения основных экспериментов были выполнены дополнительные исследования, направленные на анализ влияния архитектурных и обучающих факторов на итоговое качество. В частности, протестированы различные варианты речевых энкодеров, включая USAD, gemma-3n, Dasheng, а также комбинированные конфигурации WavLM+EATs и другие архитектурные решения. Параллельно были исследованы различные стратегии обучения, включая вариации параметрически эффективной адаптации.

Среди рассмотренных методов наилучшее качество продемонстрировал подход сброс LoRA, предполагающий повторную инициализацию адаптера перед обучением на целевой задаче. Комбинированная модель, основанная на оптимальной конфигурации энкодера и стратегии обучения, показала высокое качество в сравнении с базовой моделью, особенно в условиях обучения на ограниченных объёмах данных. Это подтверждает эффективность выбранной архитектурной и обучающей парадигмы для задач мультимодальной обработки аудио и текста при ограниченных вычислительных и датасетных ресурсах.

## Выводы

Таким образом, в рамках работы разработана гибкая и воспроизводимая мультимодальная модель, способная адаптироваться к широкому спектру акустических и языковых задач. Это было обеспечено модульной архитектурой, параметрически эффективной стратегией обучения и использованием иерархически организованного бенчмарка, позволяющего проводить систематическую и сопоставимую оценку качества.

Следует подчеркнуть, что полученные в исследовании результаты частично отличаются от выводов разработчиков бенчмарка. В качестве наиболее качественной модели-судьи по результатам проведённой оценки была определена Llama 3.3 70B Instruct, продемонстрировавшая точность 97 % на сформированном тестовом наборе. В то же время авторы Dynamic SuperB Phase 2 в своей работе отдавали предпочтение модели GPT-4o, показавшей точность 96 %.

Данное расхождение может быть обусловлено различиями в тестовых выборках, протоколах оценки или критериях агрегации метрик, что подчёркивает важность независимой валидации моделей-судей при построении комплексных оценочных пайплайнов.

**Литература**


1. Chen S. et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing //IEEE Journal of Selected Topics in Signal Processing. – 2022. – Т. 16. – №. 6. – С. 1505-1518.
2. Yang A. et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement //arXiv preprint arXiv:2409.12122. – 2024.
3. Hu E. J. et al. Lora: Low-rank adaptation of large language models //arXiv preprint arXiv:2106.09685. – 2021.