

ПРИМЕНЕНИЕ AUDIOLLM СИСТЕМЫ ДЛЯ РЕШЕНИЯ ЗАДАЧ, СВЯЗАННЫХ С ЛОКАЛИЗАЦИЕЙ ДИКТОРОВ И ДЕТЕКЦИЕЙ СПУФИНГА НА АУДИО

Никитин А.В. (ИТМО)

Научный руководитель – Митрофанов А.А. (ИТМО)

Введение. Современные системы обработки аудиосигналов играют ключевую роль в задачах безопасности и аутентификации пользователей. Одной из актуальных проблем является локализация дикторов в акустическом пространстве и выявление спуфинга (подделки голосовых данных) для предотвращения мошеннических действий. В данной работе рассматривается применение модели AudioLLM [1] для решения этих задач. Проведён обзор существующих методов, открытых наборов данных и выполнено дообучение модели для точного определения расстояния до диктора, а также локализации шумов на аудиозаписях. Особое внимание уделено созданию многоязычного набора данных для антиспуфинга [2].

Основная часть. В ходе работы были решены следующие задачи:

1. Проведён обзор существующих методов решения задач, связанных с локализацией. Анализ существующих методов решения задач, связанных с локализацией дикторов, выявил одно существенное ограничение. Так, например, в статье [3] предлагается решать задачу определения расстояния до говорящего с помощью архитектуры, основанной на рекуррентных и свёрточных нейронных сетях, обученной на наборе данных TIMIT. Однако такой экземпляр модели будет не способен решить задачу определения отношения сигнала к шуму. В связи с этим возникает необходимость в универсальной архитектуре нейронной сети, способной решать целый ряд задач.
2. Проведён обзор открытых наборов данных для решения задач локализации и обнаружения спуфинга. Многие наборы данных, такие, как TIMIT, Spatial LibriSpeech и другие, не имеют записей, содержащий естественные шумы окружающей среды, из-за чего качество модели, обученной на таком наборе данных, будет низким.
3. Сгенерирован обучающий набор данных для решения задачи локализации дикторов и локализации шумов. Сгенерированные аудиозаписи содержат эффект реверберации и различные шумы, что приближает набор данных к реальным условиям.
4. Проведено дообучение AudioLLM системы для оценки расстояния до диктора на основе акустических характеристик записи. Качество оценивалось с помощью следующих метрик: среднеквадратичная ошибка, точность. Результаты оказались сопоставимы с результатами, полученными архитектурами, способными только одну конкретную задачу.
5. Подготовлены наборы данных для решения задачи детекции спуфинга на аудио. Сгенерированы текстовые инструкции для обучения модели AudioLLM определять реальные аудиозаписи от поддельных.

Выводы. В работе предложен и протестирован подход к использованию AudioLLM для локализации дикторов и выявления спуфинга. Достигнутые результаты демонстрируют возможность применения данного метода для повышения точности и надёжности систем аудиодетекции.

Список использованных источников:

1. Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Chao Zhang. SALMONN: Towards Generic Hearing Abilities for Large Language Models // arXiv. 2023. arXiv:2310.13289. URL: <https://arxiv.org/abs/2310.13289> (дата обращения: 06.02.2025).

2. Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, Konstantin Böttinger. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset // arXiv. 2024. arXiv: 2401.09512. URL: <https://arxiv.org/abs/2401.09512> (дата обращения: 06.02.2025).
3. Michael Neri, Archontis Politis, Daniel Krause, Marco Carli, Tuomas Virtanen. Speaker Distance Estimation in Enclosures from Single-Channel Audio // arXiv, 2024. arXiv: 2403.17514. URL: <https://arxiv.org/abs/2403.17514> (дата обращения: 06.02.2025).