УДК 004.85

РАСПОЗНАВАНИЕ МЕТАДАННЫХ С ПРИМЕНЕНИЕМ СКВОЗНЫХ СИСТЕМ

Тен Л.В. (Университет ИТМО)
Научный руководитель – к.т.н. Романенко А.Н.
(Университет ИТМО)

Введение. В задаче автоматического распознавания речи (Automatic Speech Recognition, ASR) особый интерес представляет не только сама расшифровка речи, но и любые полезные свойства речевого сигнала, несущие дополнительную информацию о говорящем и/или его окружении. Эти свойства мы назовем метаданными. В это понятие входит паралингвистическая и экстралингвистическая информация, а также классификация фоновых акустических событий и сцен. Распознавание речи и метаданных в последние несколько лет осуществляется преимущественно при помощи сквозных (end-to-end) ASR-систем. Такие системы реализуются при помощи глубоких нейронных сетей и напрямую отображают последовательности акустических признаков речевого сигнала в последовательности букв или слов, что позволяет значительно упрощать обучение и дообучение моделей, а также повышает их производительность [1].

Основная часть. Распознавание различных типов метаданных, таких как пол, возраст, эмоциональное состояние говорящего и др., обычно выносится в отдельные задачи, требующие собственных наборов данных. Для построения сложных систем, способных извлекать метаданные в дополнение к расшифровке речи, используется многозадачное обучение. Его цель состоит в передаче знаний между различными задачами внутри модели с единой структурой, что улучшает ее обобщающую способность на каждой отдельной задаче. В данной работе рассматривается ряд сквозных ASR-систем, основанных на многозадачном обучении, а именно All-in-One Transformer [2], SALMONN [3], Whisper-AT [4] и Qwen-Audio [5]. Каждая из них обучена решать от двух до 30 задач на большом объеме аудиоданных, достигая на многих из них SOTA (state-of-the-art) результатов. На основе анализа указанных моделей выбраны наиболее перспективные подходы к распознаванию метаданных и проведены эксперименты для оценки точности распознавания как метаданных, так и самой речи, с использованием одних и тех же обучающих данных. Результаты моделей сравниваются между собой в рамках отдельных задач, а также с другими успешными решениями в данной области.

Выводы. Приведены результаты анализа существующих подходов к распознаванию метаданных при помощи сквозных ASR-систем. Проведены экспериментальные исследования наиболее перспективных моделей, сделаны выводы об их применимости к исследуемой задаче и предложены пути развития разрабатываемых систем.

Список использованных источников:

1. Prabhavalkar R., Hori T., Sainath T. N., Schluter R., Watanabe S. End-to-End Speech Recognition: A Survey // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2023. — Vol. 32. — Pp. 325–351.
2. Moritz N., Wichern G., Hori T., Le Roux J. All-in-One Transformer: Unifying Speech Recognition, Audio Tagging, and Event Detection // Interspeech 2020. — 2020. — Pp. 3112–3116.
3. Tang C., Yu W., Sun G., Chen X., Tan T., Li W., Lu L., Ma Z., Zhang C. SALMONN: Towards Generic Hearing Abilities for Large Language Models. — 2023. — URL: https://arxiv.org/pdf/2310.13289.pdf (дата обращения 18.02.2024).
4. Gong Y., Khurana S., Karlinsky L., Glass J. Whisper-AT: Noise-robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers. — 2023. — URL: https://arxiv.org/pdf/2307.03183.pdf (дата обращения 18.02.2024).

5. Chu Y., Xu J., Zhou X., Yang Q., Zhang S., Yan Z., Zhou C., Zhou J. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. — 2023. — URL: https://arxiv.org/pdf/2311.07919.pdf (дата обращения 18.02.2024).