УДК 004.934.1'1

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТРАТЕГИЙ ДЕКОДИРОВАНИЯ В ИНТЕГРАЛЬНЫХ СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ НА ОСНОВЕ КОНФОРМЕР-МОДЕЛИ

Капуста К.Л. (Университет ИТМО)
Научный руководитель – д.т.н. Карпов А.А (Университет ИТМО, СПб ФИЦ РАН)

Введение. Системы автоматического распознавания речи (САРР) активно развиваются в направлении улучшения качества распознавания, измеряемого показателем Word Error Rate (WER), и повышения скорости распознавания, измеряемой показателем Real Time Factor (RTF). Большинство современных САРР строятся на основе интегрального (end-to-end) подхода. Для таких систем применяются различные стратегии декодирования, однако наиболее распространёнными и эффективными с точки зрения показателей WER и RTF являются подходы нейросетевой темпоральной классификации (Connectionist Temporal Classification, CTC) [1] и нейросетевого трансдьюсера (Recurrent Neural Network Transducer, RNN-T) [2]. В данном исследовании проводится сравнение показателей WER и RTF для стратегий декодирования CTC и RNN-T в задаче распознавания русской речи при использовании современной архитектуры Fast Conformer [3].

Основная часть. Система автоматического распознавания речи может быть описана как сочетание блоков кодера, выполняющего акустическое моделирование и блока декодера, выполняющего языковое моделирование. В современных системах весь процесс выполняется в рамках единой интегральной системы, в которой все части обучаются совместно. Существует две основных стратегии преобразования скрытого состояния, получаемого в результате работы кодера, в текст – CTC и RNN-T. Подход CTC подразумевает построение прямого выравнивания между входными и выходным последовательностями токенов (символов или слов) с использованием «пустого символа». При этом подход RNN-T подразумевает наличие отдельной сети предиктора для языкового моделирования и соединительной сети, которая связывает результаты кодера и предиктора. Оба подхода имеют свои преимущества и недостатки, вытекающие из их фундаментальных особенностей. В данной работе проводится экспериментальное исследование этих подходов по показателям качества WER и скорости RTF распознавания речи. Для сравнения рассматриваемых стратегий декодирования была обучена современная интегральная модель Fast Conformer, которая хорошо себя показала в задаче распознавания речи на английском языке [1]. Используемые для обучения данные были извлечены из открытого корпуса OpenSTT [4] и охватывают разнообразные источники, такие как аудиокниги, телефонные разговоры, публичные выступления и радиопередачи.

Выводы. Результатом данной работы является сравнительный анализ подходов декодирования речи по критериям качества и скорости распознавания для задачи распознавания русской речи на основе современной интегральной конформер-модели Fast Conformer.

Список использованных источников:

1. Graves A. et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks // Proceedings of the 23rd international conference on Machine learning. – 2006. – С. 369-376.
2. Graves A. Sequence transduction with recurrent neural networks // arXiv preprint arXiv:1211.3711. – 2012.
3. Rekesh D. et al. Fast conformer with linearly scalable attention for efficient speech recognition // 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). – IEEE, 2023. – С. 1-8.
4. Slizhikova, A., Veysov, A., Nurtdinova, D., Voronin, D., Baburov, Y. Russian open speech to text (stt/asr) dataset v1.0 [Электронный ресурс]. – URL: https://github.com/snakers4/open_stt/ (дата обращения: 12.02.2024).