УДК 004.89

ТОНКАЯ НАСТРОЙКА МОДЕЛИ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ ЯЗЫКОВ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ

Алексеева Д.К. (ФГБОУ ВО «Пятигорский государственный университет»)
Научный руководитель – кандидат экономических наук, доцент Тимченко О.В.
(ФГБОУ ВО «Пятигорский государственный университет»)

Введение. Модели автоматического распознавания речи (Automatic Speech Recognition, ASR) решают задачу преобразования устной речи в текстовую форму. Проблема распознавания речи на малоресурсных языках заключается в недостатке тренировочных датасетов и сложности сбора данных, требующих привлечения экспертов в области лингвистов для их валидации. Также актуальна проблема разработки ASR-систем для бесписьменных языков. В таких случаях возможно использование международного фонетического алфавита (International Phonetic Alphabet, IPA) для создания транскрипций.

Дополнительную сложность представляет группа агглютинативных языков, в которых слова образуются путём последовательного присоединения морфем, что приводит к появлению большого количества различных словоформ. Это затрудняет работу стандартных языковых моделей и требует более гибких методов обработки.

В этих условиях для разработки ASR-систем для языков с ограниченными ресурсами эффективно использование предобученных многоязычных моделей, которые можно тонко настроить (fine-tuning) для конкретного языка с использованием небольшого количества специфических данных, а предварительное обучение на большом объёме многоязычных данных позволяет модели эффективнее адаптироваться к новым языкам благодаря переносу лингвистических паттернов между многоресурсными и малоресурсными языками [3].

Основная часть. Для решения поставленной задачи была выбрана модель Massive Multilingual Speech (MMS) [2] на базе Wav2Vec 2.0 [1] для распознавания речи на кабардинском языке. Она превосходит аналогичную ей модель XLS-R [4] по количеству данных в обучающем наборе и обеспечивает снижение вычислительных затрат благодаря использованию технологии адаптеров, позволяющих тонко настроить предобученную модель, оставляя исходные параметры без изменений.

Адаптерные слои модели MMS внедрены на каждом блоке трансформера, после последнего feed-forward слоя, и они состоят из нормализации (LayerNorm), линейных преобразований и активации ReLU. Далее языково-специфические адаптеры и выходной слой подстраиваются под уникальные характеристики языка путём дообучения на размеченных данных, что улучшает распознавание речи без необходимости создания отдельной модели для каждого языка.

Однако в большинстве случаев случайно инициализированный линейный слой, преобразующий выход модели в специфический словарь языка, не будет в полной мере соответствовать целевому языку, особенно если он относится к категории малоресурсных. Это может привести к несоответствиям в предсказаниях модели, в частности при обработке редких или сложных словоформ, что особенно актуально для агглютинативных языков, таких как кабардинский. Поэтому в процессе создания модели для распознавания речи на данном языке адаптерные слои были инициализированы заново для проведения тонкой настройки.

Также в процессе эксперимента были индивидуально подобраны гиперпараметры с учётом специфики языковых транскрипций и аудиоданных для обеспечения оптимальной точности распознавания.

В настоящее время для снижения показателей процента ошибок (Word Error Rate, WER) тестируются такие подходы, как добавление языковой модели и расширение датасета путём аугментации данных.

Выводы. В данной работе описан процесс тонкой настройки модели MMS для автоматического распознавания речи на кабардинском языке и была получена оценка качества её работы.

Список использованных источников:

1. Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in neural information processing systems. – 2020. – Т. 33. – С. 12449-12460.
2. Pratap V. et al. Scaling speech technology to 1,000+ languages //Journal of Machine Learning Research. – 2024. – Т. 25. – №. 97. – С. 1-52.
3. Protasov V. Super donors and super recipients: Studying cross-lingual transfer between high resource and low-resource languages / V., Protasov E. Stakovskii, E. Voloshina, T. Shavrina, A. Panchenko // In Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages. 2024. pp. 94-108.
4. XLS-R: selfsupervised cross-lingual speech representation learning at scale. / A. Babu [et al.] // 23rd Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 2278-2282.