УДК 004.85

ИССЛЕДОВАНИЕ МНОГОЯЗЫЧНЫХ ИНТЕГРАЛЬНЫХ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

Кутлов Р.Р. (Университет ИТМО)
Научный руководитель – к.т.н. Романенко А.Н. (ООО “ЦРТ”),
Кабаров В.И.(Университет ИТМО)

Введение. Исследования в области распознавания речи (Automatic Speech Recognition – ASR) с использованием различных моделей распознавания речи сталкиваются с несколькими проблемами при работе с аудиоданными:

1.В каждом высказывании присутствует несколько звуковых единиц, которые имеют переменную длину без явного сегментирования.

2.На этапе предварительной обработки аудиоданных отсутствует словарь входных звуковых единиц [1].

Существует несколько моделей ASR, которые позволяют найти решение этих проблем. В работе рассмотрены наиболее распространённые модели распознавания речи ASR: Wav2Vec 2.0 и HuBERT.

Основная часть. Модель ASR Wav2Vec 2.0 обучается при помощи самообучения (self-supervised). Этот способ обучения позволяет предварительно обучать модель на более доступных неразмеченных данных. Далее модель можно дообучить (fine-tuned) для конкретной цели на конкретном наборе данных [2].

Основная идея метода – это обучение модели в два этапа. Первый этап направлен на достижение наилучшего возможного представления речи и реализуется в режиме самообучения с использованием неразмеченных данных. Второй этап обучения реализуется про помощи дообучения с учителем (supervised fine-tuning), во время которой используются размеченные данные [3].

Модель ASR HuBERT обучается с использованием скрытого блока (Hidden unit) BERT (HuBERT) для self-supervised представления речи с этапом автономной кластеризации и обеспечения выравнивания целевых меток для прогнозирования потерь (prediction loss). Ключевым компонентом подхода является применение только к замаскированным областям prediction loss, что вынуждает модель по непрерывным входным данным обучаться по комбинированной акустической и языковой модели.

Основная идея метода – изучение скрытых единиц для HuBERT. Обученная на текстовых и речевых парах акустическая модель предоставляет псевдофонетические метки для каждого кадра с помощью принудительного выравнивания при обучении с частичным привлечением учителя (semi-supervised learning – SSL).

Выводы. При сравнении моделей Wav2Vec 2.0 и HuBERT, HuBERT показала наилучшую производительность при сопоставимых показателях WER.

Список использованных источников:

1. HuBERT: How to Apply BERT to Speech, Visually Explained // HuBERT: How to Apply BERT to Speech, Visually Explained [Электронный ресурс]. Режим доступа: https://jonathanbgn.com/2021/10/30/hubert-visually-explained.html (дата обращения 20.11.2022)
2. J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // – 2019
3. Speech Recognition on LibriSpeech test-clean, Papers With Code