УДК 004.8

ПОСТРОЕНИЕ ДИКТОРСКИХ ЭМБЕДДИНГОВ С ПОМОЩЬЮ ATTENTION ENCODER DECODER ДЛЯ ЗАДАЧИ ДИАРИЗАЦИИ

Аникин А. А. (ИТМО)

Научный руководитель – кандидат технических наук Новоселов С. А.
(ИТМО)

Введение. Алгоритмы дикторской диаризации на основе кластеризации широко используются в приложениях обработки речи, таких как автоматическое распознавание речи в условиях присутствия нескольких дикторов. Задача диаризации состоит в том, чтобы выделить речевые сегменты и назначить им метки дикторов. В то же время, диаризация использует модели верификации диктора, чья задача заключается в сравнении эталонной и тестовой записи.

В большинстве случаев улучшения качества диаризации можно добиться повышая качество верификации, однако, в определенных условиях эта связь перестает работать. В таких случаях необходимо аккуратнее подходить к выделению дикторских эмбеддингов. Пример нарушения данной связи можно найти в задачах конкурса CHiME-8. В данной работе будет представлена новая модель верификации диктора на основе Attention Encoder Decoder (AED) и алгоритм диаризации на основе кластеризации для конкурса CHiME-8.

Основная часть. CHiME-8 — это конкурс посвященный автоматическому распознаванию речи в условиях присутствия нескольких дикторов. Главное отличие конкурса CHiME-8 от предыдущих лет состоит в увеличении максимального числа дикторов, которые могут присутствовать в записи. В связи с этим отдельное внимание было уделено алгоритму оценки числа дикторов, который включает в себя два шага фильтрации сегментов со смешанной речью и маленьких кластеров.

Для определения сегментов со смешанной речью была обучена модель для выделения сразу двух эмбеддингов диктора из короткого сегмента. Эта модель состоит из кодировщика — wav2vec XLSR_53 [1] и декодировщика — 12-слойный трансформер. Процесс обучения модели похож на обучение Whisper [2]. Модель предсказывает токены начала и конца сегмента и затем предсказывает метку диктора, уникальную в пределах всей обучающей выборки. В качестве функции ошибки для токенов времени использовалась Cross Entropy и AAM-Softmax для меток диктора.

Для обучения модели выделения дикторских эмбеддингов использовались синтетические склейки двух дикторов из VoxCeleb2. Для замера качества верификации использовались тестовые протоколы VoxCeleb1-O и VOiCES. Для подсчета диаризационных метрик использовались наборы данных из CHiME-8: DiPCo, CHiME, Mixer6, NOTSOFAR.

Выводы. Использование новой модели AED повысило точность оценки числа дикторов за счет возможности фильтрации речевых сегментов с пересекающейся речью и позволило работать с записями содержащими большее количество дикторов (до восьми). Новый алгоритм диаризации показал лучшее качество диаризации в сравнении с решением для CHiME-7.

Список использованных источников:

1. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M. (2021) Unsupervised Cross-Lingual Representation Learning for Speech Recognition. Proc. Interspeech 2021, 2426-2430, doi: 10.21437/Interspeech.2021-329
2. Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision."International conference on machine learning. PMLR, 2023.28492-28518