УДК 004.891.2

ИСПОЛЬЗОВАНИЕ АРХИТЕКТУРЫ TRANSFORMER В ЗАДАЧЕ ДИАРИЗАЦИИ

Авдеева А.С. (Университет ИТМО)
Научный руководитель – к.т.н. Новоселов С.А.
(ООО “ЦРТ-Инновации”)

Введение. Диаризация – это процесс разметки речевой аудиозаписи, который заключается в сопоставлении каждому гомогенному сегменту речи идентификатора диктора. Задача диаризации редко рассматривается отдельно и обычно возникает при реализации различных алгоритмов распознавания диктора или речи. Поэтому развитие диаризации во многом связано с развитием смежных областей. Широко распространен подход, основанный на использовании дикторских моделей, построенных на сегментах речи, в качестве входных признаков для диаризационных алгоритмов. Данное исследование рассматривает влияние качества и условий обучения различных дикторских энкодеров для дальнейшего применения в алгоритме диаризации.

Основная часть. Существующие методы диаризации можно разделить на end-to-end подходы и каскадные алгоритмы. Каскадный алгоритм обычно состоит из нескольких этапов: предварительная обработка, применение детектора речи, сегментация, извлечение признаков и кластеризация. Один из наиболее простых подходов среди каскадных алгоритмов основан на сегментации с некоторым постоянным окном, извлечении из каждого такого сегмента дикторской модели с помощью предобученного энкодера и последующей кластеризации [1], [2]. Область распознавания диктора непрерывно развивается в настоящее время, множество различных архитектур нейронных сетей успешно применяется в этой задаче. Однако, вследствие специфики постановки задачи диаризации, возникает необходимость обеспечивать построение устойчивой дикторской модели на коротком речевом сегменте, чтодо сих пор является сложной задачей [3]. Поэтому дикторские энкодеры, показывающие хорошее качество в рамках идентификационной метрики могут не до конца удовлетворять условиям задачи диаризации. Самообучающиеся модели на основе архитектуры Transformer [4], недавно показавшие хорошее качество в различных задачах обработки речевых данных – распознавания речи [4], диктора и языка [5], могут быть также использованы в задаче диаризации. Использование такой модели для построения дикторских представлений и последующей кластеризации алгоритмом AHC позволяет получить лучшее качество, чем приосуществлении кластеризации на представлениях из других дикторских энкодеров, таких как ECAPA-TDNN [6], TDNN [2], ResNet [7].

Выводы. Предложен алгоритм использования представлений из архитектуры Transformer в задаче диаризации. Алгоритм на основе подобного дикторского энкодера показывает лучшее качество, чем на других широко известных энкодерах, таких как TDNN, ECAPA-TDNN, ResNet.

Список использованных источников:

1. Sell, Gregory and Daniel Garcia-Romero. “Speaker diarization with plda i-vectorscoring and unsupervised calibration.” 2014 IEEE Spoken Language Technology Workshop (SLT) (2014): 413-417.
2. Rouvier, Mickael, Pierre-Michel Bousquet and Benoit Favre. “Speaker diarizationthrough speaker embeddings.” 2015 23rd European Signal Processing Conference (EUSIPCO) (2015): 2082-2086.

3. Novoselov, Sergey, Vladimir Volokhov and Galina Lavrentyeva. “Universal speakerrecognition encoders for different speech segments duration.” ArXiv abs/2210.16231 (2022): n. pag.
4. Schneider, Steffen, Alexei Baevski, Ronan Collobert and Michael Auli. “wav2vec: Unsupervised Pre-training for Speech Recognition.” Interspeech (2019).
5. Fan, Zhiyun, Meng Li, Shiyu Zhou and Bo Xu. “Exploring wav2vec 2.0 on speakerverification and language identification.” Interspeech (2020).
6. Dawalatabad, Nauman, Mirco Ravanelli, Franccois Grondin, Jenthe Thienpondt, Brecht Desplanques and Hwidong Na. “ECAPA-TDNN Embeddings for Speaker Diarization.” Interspeech(2021).
7. Jakubec, Maros, Eva Lieskovská and Roman Jarina. “Speaker Recognition withResNet and VGG Networks.” 2021 31st International Conference Radioelektronika
(RADIOELEKTRONIKA) (2021): 1-5.