УДК 004.934

ОБЗОР END-TO-END МЕТОДОВ РАЗДЕЛЕНИЯ РЕЧИ ДИКТОРОВ

Капранова В.О. (Университет ИТМО**), PhD Астапов С.С. (**Университет ИТМО**),**
Кабаров В.И. (Университет ИТМО**)**
Научный руководитель – Попов Д.В.
(ООО ЦРТ-Инновации**)**

В работе рассматриваются методы диаризации, объединяющие в себе классические и end-to-end подходы.

Введение. В современном мире очень востребованы системы, позволяющие разделить реплики нескольких дикторов из аудиозаписей. Такая задача осложняется тем, что в разговоре нескольких людей довольно часто встречаются наложения реплик. Для решения этой задачи были разработаны методы, такие как сквозная нейронная диаризация (EEND) и методы на основе предварительно вычисленных характеристик интересующего говорящего (Target-Speaker ASR, Speaker Beam, Voice Filter, GSS и другие).

Основная часть.
Текущие методы разделения реплик дикторов в большинстве своём основаны на кластеризации эмбеддингов дикторов, извлекаемых из участков аудиозаписи. В то время как некоторые методы обеспечивают контролируемую кластеризацию эмбеддингов дикторов, наиболее распространенным подходом является кластеризация x-векторов без учителя. Поскольку наивная кластеризация x-вектора приводит к плохой производительности, были предложены различные методы для улучшения производительности, например, повторная оценка вероятностного линейного дискриминантного анализа (PLDA) и пересегментация скрытой марковской модели (HMM) в вариационной байесовской модели (VB). С точки зрения обработки с перекрытием, большинство методов сначала обнаруживают перекрывающиеся кадры, а затем назначают второго диктора для обнаруженных кадров на основе эвристики или результатов пересегментации VB.
Другое направление основано на кластеризации перекрывающихся сегментов. Сначала извлекаются перекрывающиеся сегменты, используя нейронную сеть, предсказывающую регионы с перекрытием речи, а затем применяет кластеризацию для эмбеддингов, извлеченных из каждого из них. Но точность такого метода не превосходит точность нейросетевых end-to-end методов.
Следующий подход называется EEND. Суть данных методов в том, чтобы рассчитать активность нескольких дикторов, каждая из которых соответствует одному говорящему. Последние модели могут выводить гибкое количество активных дикторов с помощью модулей вычисления аттракторов на основе кодировщика-декодера (EDA) или EEND с условием говорящего (SC-EEND).
В работе рассматриваются два подхода, объединяющие в себе преимущества диаризации на основе классических и end-to-end методов.

Выводы. В рамках исследования был произведён обзор методов, объединяющих в себе классическую и end-to-end диаризацию. Применение таких алгоритмов позволит повысить качество систем разделения реплик дикторов, сочетая сильные стороны двух основных подходов к диаризации.