УДК 004.93

РЕАЛИЗАЦИЯ МОДЕЛИ ДЛЯ ПОДСЧЕТА ЧИСЛА ДИКТОРОВ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Евсеева Е.С. (Университет ИТМО)
Научный руководитель – с.н.с. Астапов С.С.
(Университет ИТМО)

В данной работе рассматривается задача подсчета числа активных дикторов и описывается ее решение на основе рекуррентных нейронных сетей.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №620172 «Определение структуры полилога с применением лексических и нелексических признаков речи нескольких дикторов».

Оценка максимального числа одновременно говорящих дикторов тесно связана с проблемой их идентификации в задаче диаризации дикторов. Диаризация дикторов, – процесс разделения входящего аудиопотока на однородные сегменты в соответствии с принадлежностью аудиопотока тому или иному диктору, по-прежнему остается трудной задачей для приложений, используемых в присутствии нескольких активных дикторов. Когда источники полностью перекрываются, сегментация и идентификация диктора затруднены. В этом случае диаризацию дикторов можно улучшить, пометив каждый входной кадр количеством активных дикторов. Для этого аудиозапись нарезается на фрагменты определенной длины и для каждого из них определяется число активных дикторов на записи.

Для решения данной задачи была определена архитектура на основе рекуррентных нейронных сетей с использованием LSTM. Данный подкласс нейронных сетей успешно используется для анализа предыдущих состояний в таких задачах как распознавание речи или обработка естественного языка, и используется в данной работе для обработки длинных фрагментов сигнала, например, 5-секундных кадров. Для обучения модели был создан набор данных смесей дикторов на основе открытого корпуса LibriSpeech с речью различных дикторов. На вход нейронной сети подавались фрагменты сигнала различной длительности, а также различные признаки сигнала, такие как частотный спектр, мел-спектрограмма, MFCC.

В данной работе была рассмотрена реализация модели для подсчета числа активных дикторов на основе рекуррентных нейронных сетей и приведено ее сравнение с готовым решением на открытом корпусе AMI с помощью метрик Recall, Precision и F1-score. Рассмотренная модель показала улучшение точности определения числа дикторов. Однако, для практического применения модель следует улучшить, уменьшив длительность принимаемых фрагментов аудио. Это позволит более точно определять временные границы активности нескольких дикторов и облегчит их диаризацию.