УДК 004.934

Диаризация двухканальных аудиозаписей для задачи распознавания речи.

А.Е. Гусев, А.С. Авдеева
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – д. т. н., профессор, Ю.Н. Матвеев
(Университет ИТМО, г. Санкт-Петербург)

Одной из важных задач при работе с аудиоданными является разделение дикторов на фонограмме - диаризация - процесс разделения входящего аудиопотока на однородные сегменты в соответствии с принадлежностью аудиопотока тому или иному говорящему. Диаризация может применяться в задачах автоматического распознавания речи для сопоставления распознанной речи с говорящим, поиска речевых фрагментов, сказанных определённым диктором (при дополнительном использование модели диктора), простого по-дикторного разделения аудио записей, определения числа дикторов, говорящих на одной аудио записи, и др [1].

В основе диаризационой системы могут лежать алгоритмы, использующие скрытые марковские модели, SVM классификаторы, алгоритмы на основе гауссовых смесей, i-вектора [2] и т.д В последние несколько лет набирают популярность диаризационные системы построенные на основе LSTM и TDNN [3] сетей, что позволило уменьшить значение DER при работе подобных систем. Несмотря на это, качества работы диаризационных систем всё ещё недостаточно для решения многих практических задач обработки аудио потока.

Другим способом решения проблемы разделения одной аудизаписи на несколько по дикторных фрагментов является применения микрофонных решёток, позволяющих производить одновременную запись источника звука на несколько пространственно-разнесённых микрофона. Далее благодаря алгоритмам, наподобие DUET, возможно произвести разделение аудизаписи по источникам звука, в том числе и дикторам. Недостатком подобного способа является его низкая помехоустойчивость и необходимость применять сложные и дорогостоящие микрофонные решётки для достижения хорошего качества работы алгоритма.

В данной работе рассмотрен способ по-дикторного разделения аудио данных, полученных с простой двух микрофонной решётки. Применение диаризационной системы, использующей фазовую информацию аудио записей совместно с дикторными x-векторами позволили уменьшить значение word DER с 18.8% при использование отдельно фаз, 16% при использование отдельно x-векторов до 13%. Кроме того, рассмотрены несколько методов, позволяющих уменьшить дикторную составляющую DER за счёт оценки качества работы алгоритма для различных слов, что позволяет точнее сопоставлять речь с говорящим за счёт отбрасывания речевых участков с низким качеством работы диаризационной системы.

Литература

1. Beigi H. Fundamentals of Speaker Recognition. SpringerLink : Bucher. — Springer, 2011. — ISBN: 9780387775920.
2. Кудашев О. Ю. Агломеративная кластеризация речевых сегментов фонограммы на основе байесовского информационного критерия // Научнотехнический вестник информационных технологий, механики и оптики. —2013. — № 1. — С. 90–93.
3. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, X-vectors: Robust DNNembeddings for speaker recognition, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). IEEE, 2018, pp. 5329–5333.