Факультет вычислительной математики и кибернетики

МГУ имени М.В. Ломоносова

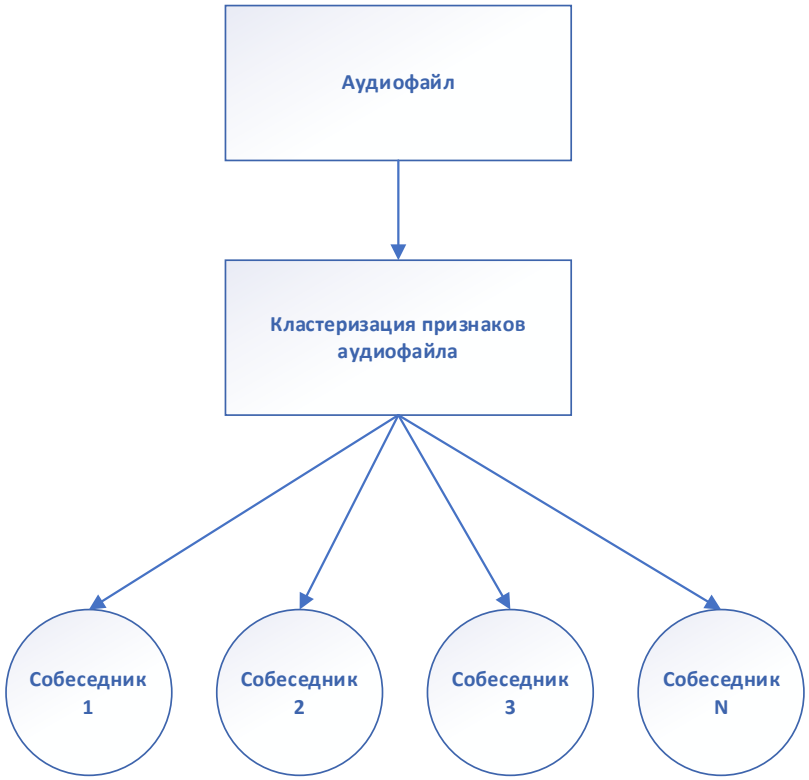
Программа профессиональной переподготовки «Разработчик компьютерных технологий»

Тема: «Решение задачи диаризации методами машинного обучения»

Синев Олег Сергеевич
Смирнов Илья Николаевич

Постановка задачи

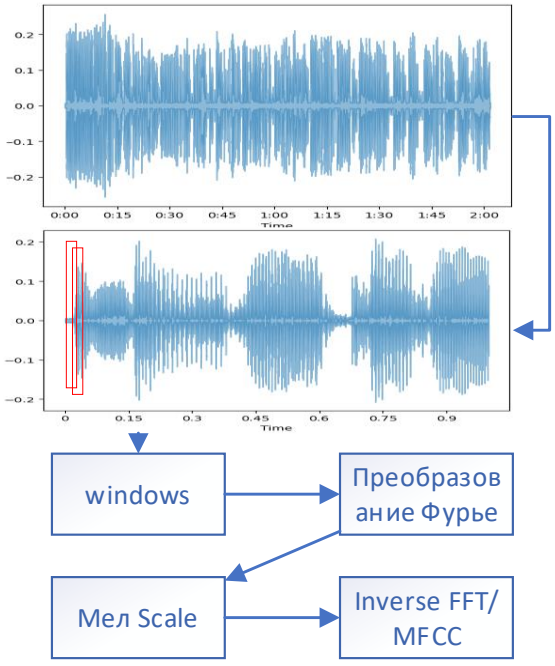
Цель работы – разработать алгоритм, автоматически распознающий количество собеседников в аудиозаписи или, иными словами, диаризацию (Speaker Diarization — SD, которая известна в англоязычной литературе как Who Spoke When). На вход подаётся аудиофайл с диалогом нескольких человек, на выход выдаётся количество собеседников, обнаруженных в аудиозаписи.

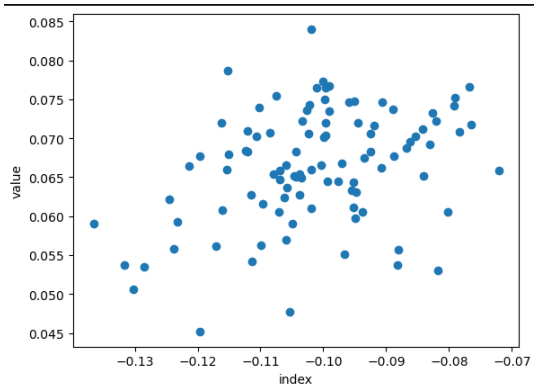


Постановка задачи

Для достижения цели были поставлены следующие задачи:

- Удаление из аудиозаписи посторонних шумов и тишины.(librosa.effects.split, noisereduce)
- Разбиение аудиозаписи на фреймы по 1 секунде
- Выделение признаков каждого фрейма.(librosa.feature.mfcc)
- Обучить модель KMeans на наборе признаков с разным набором кластеров.
- Протестировать алгоритм на реальных данных.





Актуальность работы

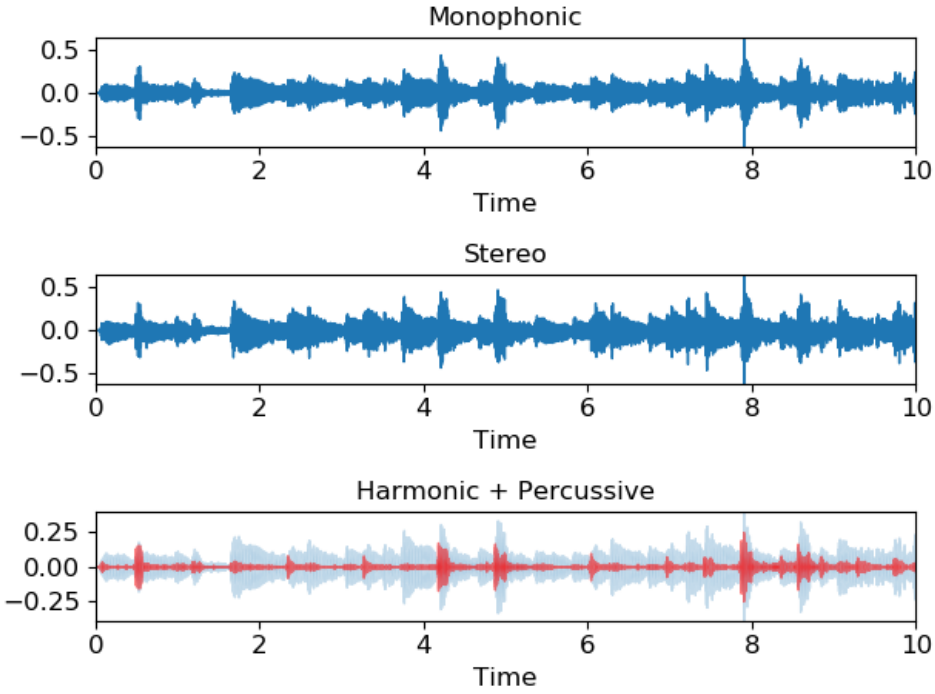
Результаты данной работыя востребованы в биометрическом поиске, голосовой верификации, разграничении прав доступа к информации, создании субтитров к видеозаписям, системах умного дома, криминалистике в качестве помощи к системам распознавания речи и т.д.

Подходы к реализации задачи

Что такое librosa?

Librosa — это пакет Python для анализа и синтеза аудио сигналов.

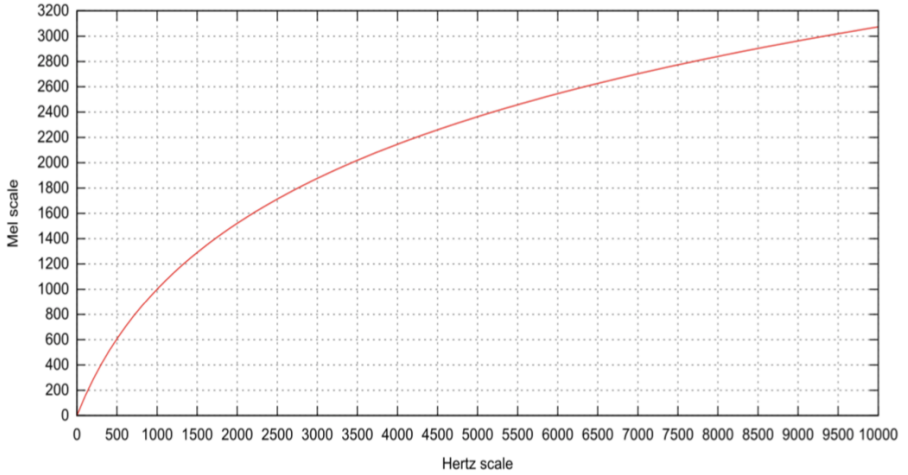
Он предоставляет строительные блоки для создания структур, которые помогают получать информацию о музыке и человеческой речи в аудиозаписи.



Подходы к реализации задачи

Librosa предоставляет возможности извлечения Мел-Кепстральных коэффициентов (MFCC) из аудиофайла. Шкала мел описывает отношение высоты чистого тона (мел) к фактической измеренной частоте (Гц).

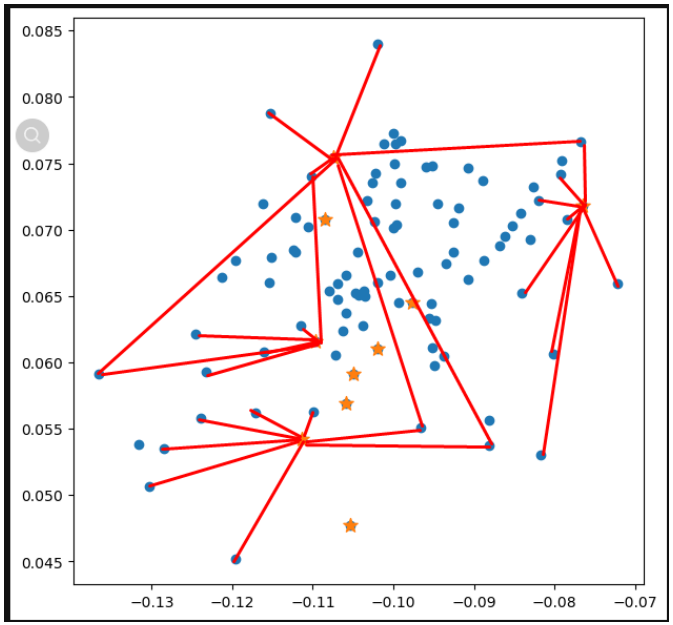
$$mel = 1127.01048 \ln(1 + \frac{freq}{700})$$



Подходы к реализации задачи

Алгоритм K-средних. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

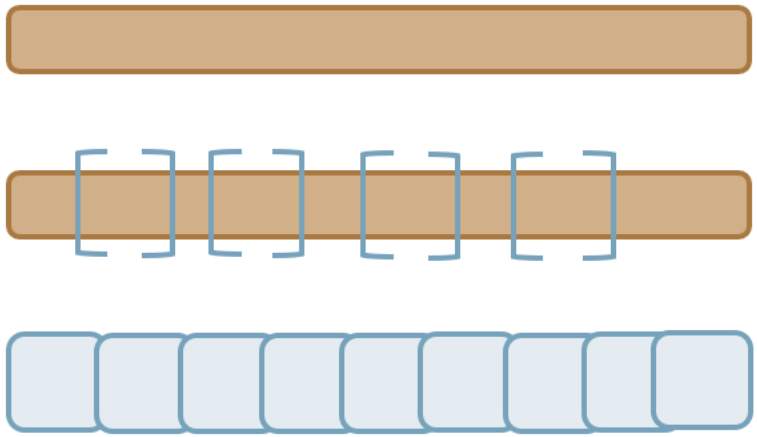
$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2$$



$$\sqrt{\sum (x_1 - x_2)^2}$$

Извлечение признаков из аудиозаписи

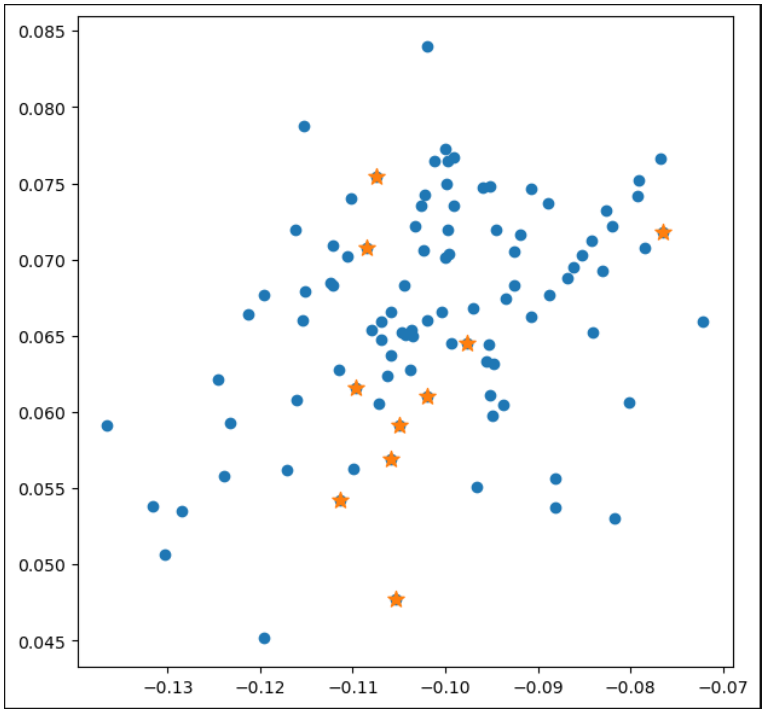
Первый этап — извлечение характеристик. Предварительно аудиофайл разделяется на фреймы фиксированной длины с небольшим наложением, далее из полученных фреймов получаем мел-кепстральные коэффициенты с помощью библиотеки Librosa.



Метод К-средник

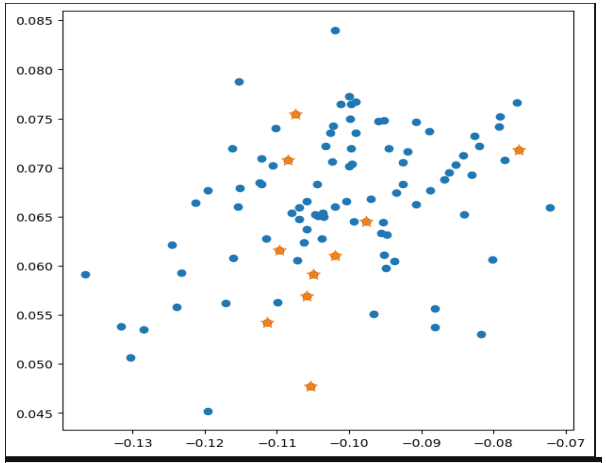
На следующем этапе из полученных данных строим точки. Выбрав случайную точку обозначаем её как центройд средствами matplotlib.plot.

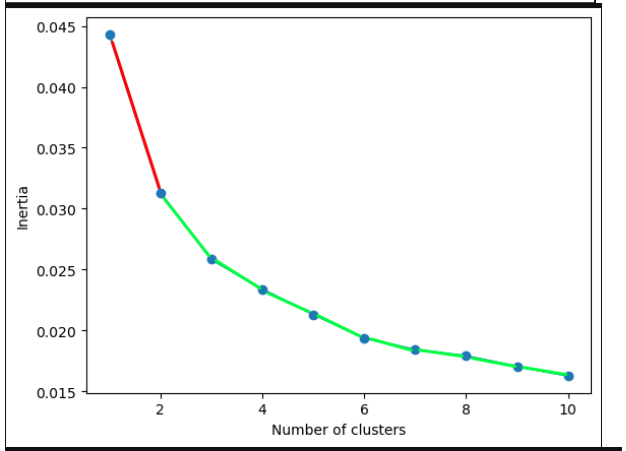
Центройды являются центрами будущих кластеров.



Кластеризация по методу локтя

С помощью метода Локтя строим модель KMeans, с определением оптимального количества кластеров. Для ограничения среднеквадратичного расстояния между кластерами добавляем пороговое значение для алгоритма.

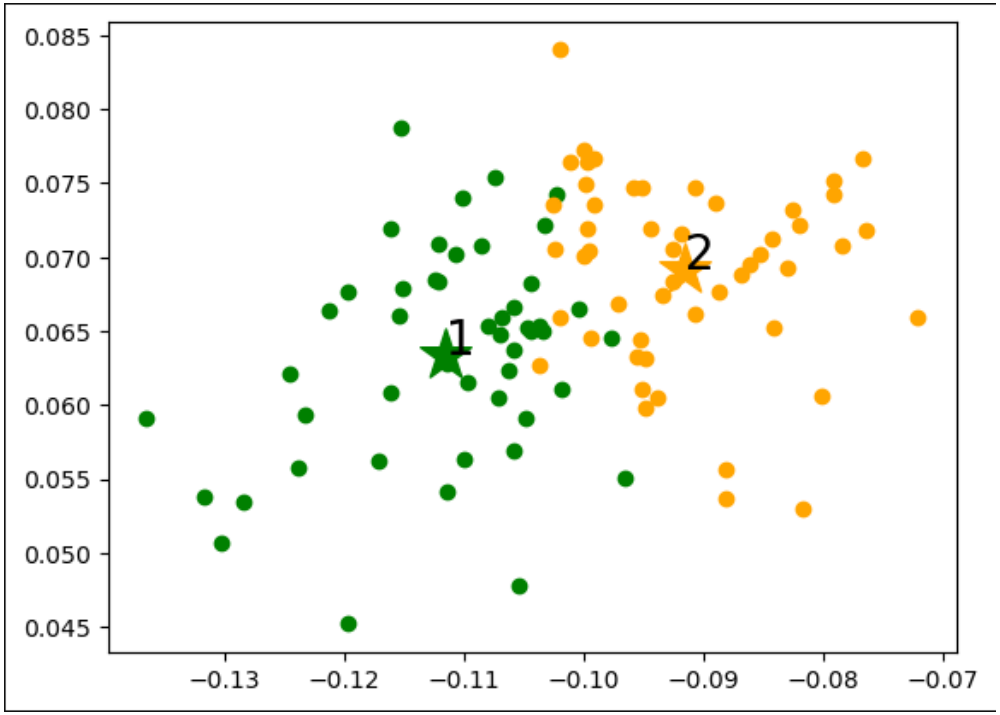




$$k_{opt} = argmin(\frac{inertia_{k+1} - inertia_k}{inertia_k - inertia_{k-1}})$$

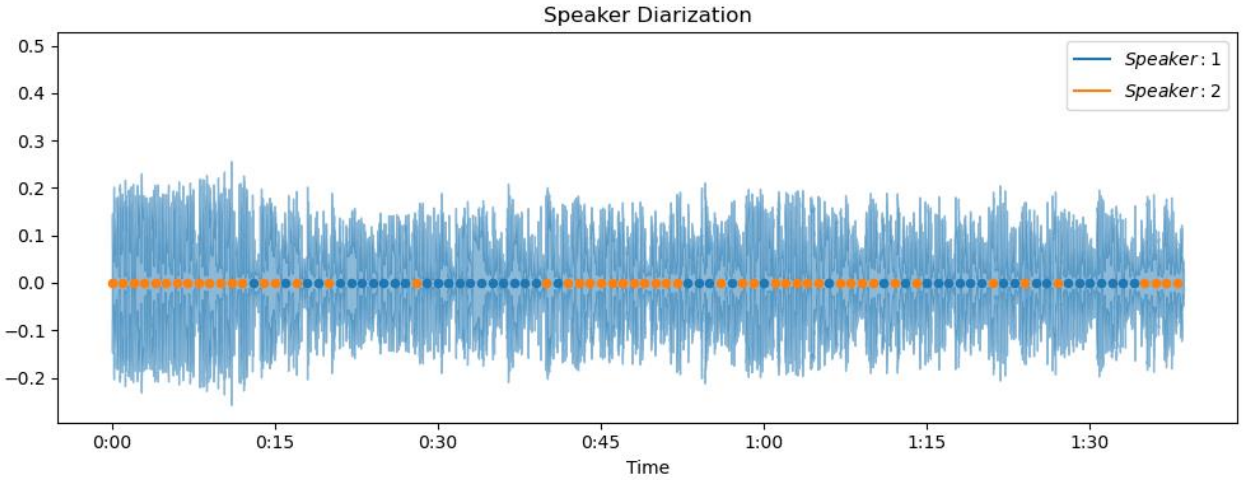
Кластеризация данных

Кластеризуем набор признаков аудиозаписи на обученной модели KMeans. В результате получаем количество кластеров, равное количеству собеседников в аудиозаписи.

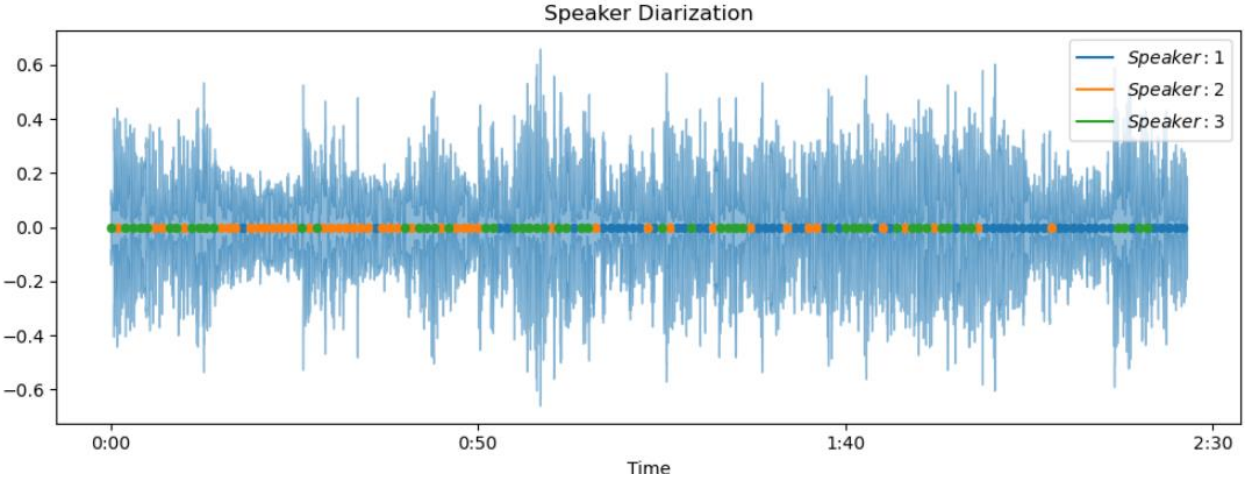


Результаты работы

Пример реализации на записи с двумя голосами



Пример реализации на записи с тремя голосами



Выводы

- В случае уменьшения длительности фреймов качество определения голоса возрастает
- Зависимость от параметров анализа данных: наложение, длительность сегментов, пороговые значения
- Относительная несложность реализации задачи
- Реализация актуальна вне зависимости от языка речи