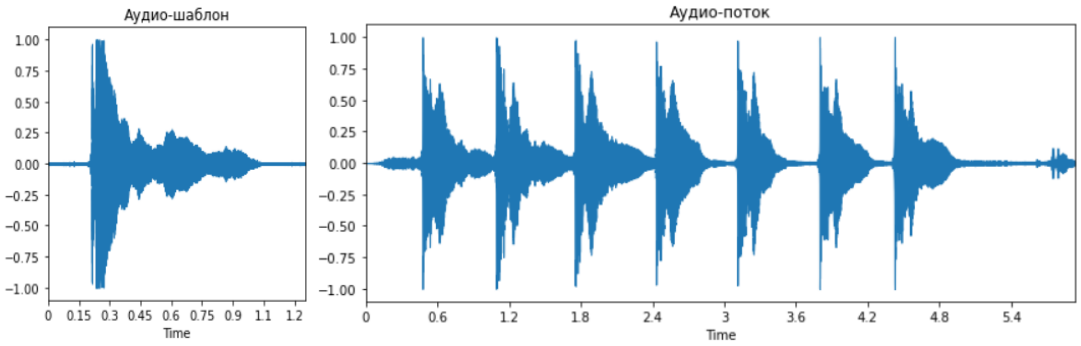
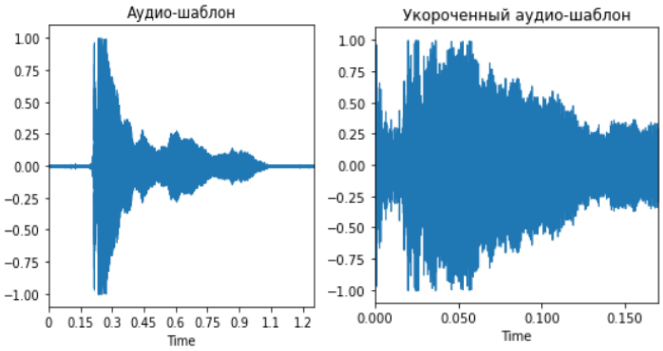
Описание метода

Обнаружение участков аудио-потока схожего по звучанию с шаблоном производится за счет использования скользящего окна вдоль временной оси.



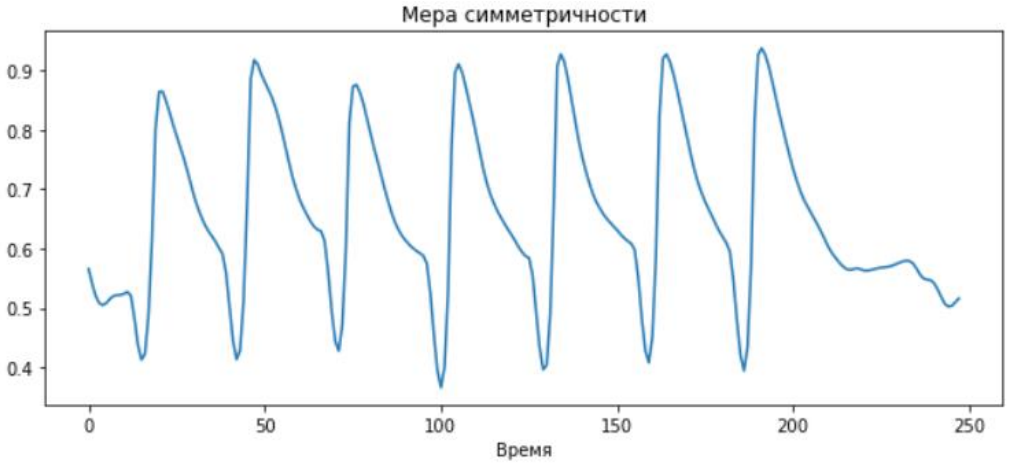
Перед сравнение двух аудиозаписей обрезается шаблон, так как помимо основного звука он также содержит и участки тишины в начале и в конце. Для этого отбираются значения, превышающие по модулю квантиль порядка *cut-quantile (задаваемый параметр, по умолчанию равен 0.96)* распределения абсолютных значений цифрового представления аудио-шаблона. Участок шаблона до первого выбранного значения, а также участок после последнего значения считаются участками тишины.



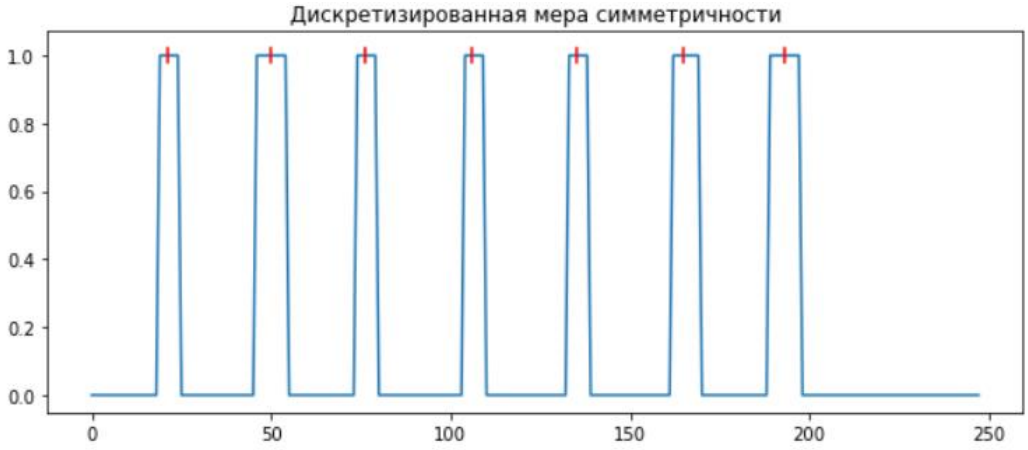
После этого шаблон и аудио-поток преобразовываются в признаки с помощью функции librosa.feature.mfcc, которая возвращает двумерный массив (n_mfcc x time). Параметр *n_mfcc* задаваемый параметр, по умолчанию равен 20. Массив признаков шаблона проходится вдоль массива аудио-потока как окно. Мера схожести между признаками шаблона и потока вычисляется как поэлементное произведение двух матриц с последующим вычислением суммы всех произведений. Непосредственно до перемножения обе матрицы нормируются (делятся на норму вектора, получаемого из соответствующей матрицы выписыванием подряд всех строк). Таким образом, мера схожести есть не что иное, как косинус угла между двумя признаковыми векторами и чем ближе это значение к 1, тем более схожи эти вектора в признаковом пространстве, а соответственно и их матрицы тоже. Однако, чтобы не

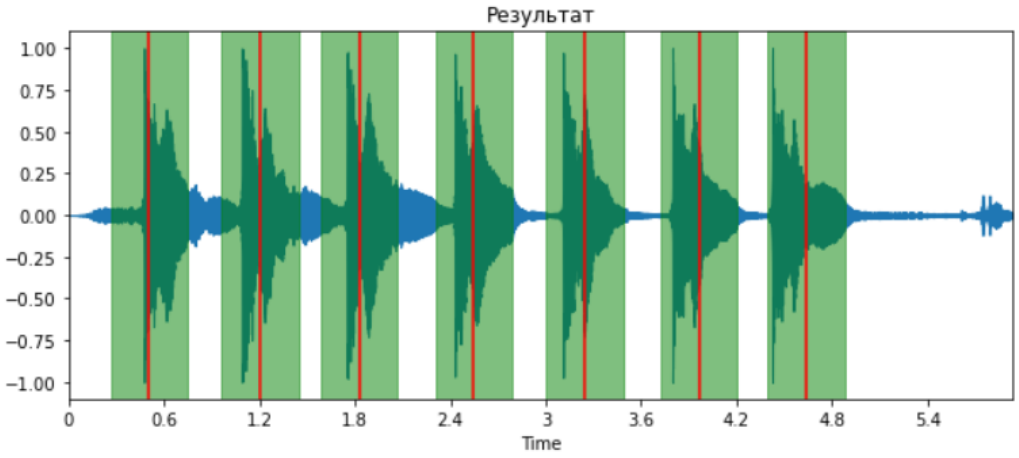
включить в результат ложные шумовые участки схожие по форме, до нормировки проводится сравнение максимального значения участка потока внутри окна и медианного значения шаблона.

Участки аудио-потока, которые имеют сильную схожесть с шаблоном, сохраняют ее и при небольших сдвигах вправо и влево. Отсюда вытекает тот факт, что график меры схожести получается гладким, как на рисунке ниже.



Чтобы получить участки локальных максимумов, мера схожести дискретизируется с помощью сравнения с пороговым значением (*threshold, задаваемый параметр, по умолчанию равен 0.8)*. В участках полученной функции, где она имеет значение 1, осуществляется поиск середины, которая далее выступает как точка совпадения аудио-потока и шаблона. Недостатком такого подхода может быть смещенность найденной точки, так как она по сути является грубым приближением локального максимума.





Данный реализация не применима, в случае если аудио-шаблон сильно отличается от звуков в потоке или полностью отличается, так как возможны ложные срабатывания.