Добрый день, меня зовут Баранова Дарья. Тема моей работы – Разработка системы идентификатора говорящего.

Пояснительная записка содержит…...

Целью работы является разработка системы идентификации диктора по голосу.

Задача идентификации говорящего на основе биометрии становится все более актуальной и находит широкое применение в реальных задачах. Несмотря на существование на рынке готовых решений, их алгоритм работы неизвестен, код закрыт, метод не описан. Востребованность на рынке продолжает увеличиваться и исследование области голосовой биометрии исследуется более активно по сравнению с другими видами биометрии. Например, она используется для подтверждения личности на горячих линиях и при звонках в call-центры, при голосовой аутентификации в личных интеллектуальных устройствах, таких как сотовые телефоны, автомобили и ноутбуки. Такая идентификация не только гарантирует безопасность транзакций банковской торговли и удаленных платежей, но также автоматизирует ряд задач, ведет к сокращению времени и росту лояльности. Биометрическая идентификация на основе голоса также может широко применяться в судебной практике для расследования подозреваемого на предмет признания его виновным, для наблюдения и автоматической идентификации личности на расстоянии (возможно, без ведома самой личности). Она также может выступать частью задачи автоматического определения количества говорящих в записанном аудиофрагменте и сегментации фрагмента на части, в которых звучит голос определенного диктора.

Задача заключается в создании системы идентификации личности диктора по его голосу. Система должна иметь возможность сохранять голос диктора в своей базе и выполнять идентификацию по запросу. Идентификация будет происходить по закрытому множеству сохраненных голосов – в этом случае ответ всегда будет существовать и диктор всегда будет найден. Также необходимо, чтобы на ответ системы влиял только голос диктора и никак не влиял произнесенный на аудиозаписи текст – система будет являться текстонезависимой.

Чтобы извлечь признаки из полученного аудиосигнала будем использовать мел-кепстральные коэффициенты. Используется именно это представление данных так как оно позволяет учитывать человеческие особенности восприятия звука

Исходный сигнал нарезают на перекрывающиеся фреймы (окна) небольшой длины (10-40 мс). Затем к получившимся фреймам применяют окно Хемминга (чтобы сгладить значения на границах фреймов и уменьшить утечку спектра), делают быстрое преобразование Фурье (FFT):

𝑋[𝑘]= ∑𝑥[𝑛]∗𝑒−2∗𝜋∗𝑖∗𝑘∗𝑛/𝑁𝑁−1𝑛=0,0≤𝑘<𝑁 (1)

и получают спектральную плотность мощности (распределение мощности сигнала в зависимости от частоты, то есть мощность, приходящаяся на единичный интервал частоты).

Затем специальной «гребёнкой» фильтров, расположенных равномерно по мел-шкале (рисунок ниже) делают мел-спектрограмму (каждый мел-фильтр - это треугольная оконная функция, которая суммирует количество энергии на определённом диапазоне частот и тем самым дает мел-коэффициент), после чего логарифмируют полученные результаты (считается, что таким образом понижается чувствительность коэффициентов к шумам) и применяют дискретное косинусное преобразование (DCT) —алгоритм сжатия данных.

задача распознавания диктора будет решаться через построение оптимальных для каждого диктора моделей гауссовской смеси распределений (Gaussian Mixture Model)

- это функция, состоящая из нескольких гауссианов. Другими словами, она (Модель Гауссовой смеси) представляет собой смесь многомерных гауссовых распределений вероятностей (многомерных нормальных распределений), которые наилучшим образом моделируют входной набор данных.

GMM в нашем случае она будет описывать распределение данных и представлять оценку плотности. Голубые окружности показывают смоделированное GMM распределение входных данных.

UBM (Universal Background Model) также является большой моделью гауссовой смеси (GMM), обученной для представления дикторонезависимого распределения признаков

Для оценки точности распознавания существует показатели точности ……….

𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а 𝑟𝑒𝑐𝑎𝑙𝑙 показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм. Введение 𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 не позволяет нам записывать все объекты в один класс, так как в этом случае мы получаем рост уровня False Positive (FP). 𝑅𝑒𝑐𝑎𝑙𝑙 демонстрирует способность алгоритма обнаруживать данный класс вообще, а 𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 — способность отличать этот класс от других классов.

Разработка велась на Python 3 в данных средах разработки. Использовались данные библиотеки для удобства разработки.

Таким образом, описанный алгоритм обучения и построения модели GMM-UBM был опробован с различными параметрами:

1. Количество MFCC: 8, 13, 20

2. Тип ковариации в GMM: full, diag, tied

3. Количество компонент: 8, 16, 32

4. С разной длительностью тестовых данных: 5 сек, 20 сек, 40 сек

Для обучения GMM брались аудиофрагменты речи длительностью 20 секунд

Таким образом, поставленная цель работы была достигнута и была разработана система идентификации говорящего. В процессе работы решены следующие задачи. Достигнутая степень успешного распознавания составляет более 99%