**Исследование возможности применения двумерной свертки в задаче распознавания эмоций**

Тулкит OpenSMILE позволяет извлекать большое количество параметров из звукового сигнала. На данный момент, мы используем набор параметров eGeMAPS, который содержит 88 параметров, которые являются различными статистическими функциями, примененными к 23 низкоуровневым дескрипторам.

ОЧЕНЬ упрощенно (насколько мне хватило знаний разобраться), процесс извлечения признаков по умолчанию выглядит следующим образом:

1. Получить звуковой сигнал на вход
2. Фрейминг (плавающее окно (Хэмминга?) длиной ?мс с перекрытием ?мс)
3. Вычисление 23 низкоуровневых дескрипторов для каждого фрейма, со сглаживанием по 3 фрейма
4. Применение статистических функций к низкоуровневым дескрипторам на всех полученных фреймах (для некоторых – только на фреймах, в которых значение вычисленной фундаментальной частоты ≠ 0 (игнорирование тишины)
5. Запись 88 параметров в файл

OpenSMILE – очень гибкий тулкит, в котором, по заявлению авторов, кастомизируется абсолютно все, начиная от применяемой оконной функции и её параметров, и заканчивая кастомизацией процесса извлечения всех доступных низкоуровневых дескрипторов. Однако, документация для этого тулкита оставляет желать лучшего и до сих пор (с 2016 года) значится как work in progress. Поэтому, я не смог найти не только информации, как изменить некоторые параметры, но даже не смог понять, какие параметры используются по умолчанию (вопросительные знаки в пункте 2). В данный момент пытаюсь связаться с разработчиками через StackOverflow.

Однако, достаточно простым образом (с помощью ключа в командной строке или изменением конфигурационного файла) можно внести изменения в процесс извлечения признаков и получить для каждого входного файла не 88 чисел, а гораздо больше. Здесь есть два возможных пути, ведущих к немного разным результатам:

1. Ключ -appendlldscv. При применении этого ключа, в процессе извлечения признаков полностью пропускается шаг 4, и все низкоуровневые признаки для каждого фрейма пишутся в итоговый файл. Получаем не вектор размерности 88, а матрицу размерности MxN, где M – количество фреймов во входном файле, N – количество низкоуровневых дескрипторов. Очевидно, что M зависит от длительности файла и параметров оконной функции. На текущий момент, я не знаю как изменить эти параметры. N зависит от выбранного набора признаков и соответствующего конфигурационного файла. В случае eGeMAPS N = 23, самый широкий (и, судя по всему, максимальный) набор предоставляет набор ComParE конференции INTERSPEECH 2015, N = 34.
2. Изменение конфигурационного файла *FrameModeFunctionals.conf.inc*. Этот файл отвечает за то, на каком временном промежутке вычисляются статистические функции от низкоуровневых дескрипторов на шаге 4 процесса. По умолчанию, они вычисляются на всей длине сигнала: frameMode = full. Доступны следующие другие варианты:
   1. Вычисление статистических функций на нескольких плавающих отрезках фиксированной длительности и ширины шага (в секундах):

frameMode = fixed

frameSize = 5 // длительность

frameStep = 2 // ширина шага

* 1. Вычисление функций на заданном списке временных сегментов (к примеру, 1сек-2сек, 3сек-3.5сек, 3.5сек–конец файла):

frameMode = list

frameList = 1s-2s,3s-3.5s,3.5s-E

В результате извлечения признаков с такими параметрами, мы получим матрицу MxN, где M – количество отрезков, N – количество параметров в выбранном наборе. Доступно множество вариантов: eGeMAPS (88), INTERSPEECH 2009 (384), INTERSPEECH 2010 (1582), openSMILE emotion feature set (6552).

Возможность извлечения не одномерного вектора признаков, а матрицы признаков, обеспечивает возможность применения сверточных глубоких нейронных сетей. Однако, такие нейросети целесообразно применять, только если обрабатываемые данные имеют какую-то связную структуру. И если в случае с одномерной сверткой очевидно наличие такой структуры, а сверточные нейронные сети уже применяются к сырому звуковому сигналу в задачах распознавания эмоций, возможность и целесообразность применения двумерной свертки – открытый вопрос. В связи с этим, принято решение провести небольшой экспериментальный анализ данных, полученных описанными выше способами на предмет наличия какой-либо структуры.

Методика экспериментов:

1. Извлечение матрицы признаков eGeMAPS из файла 03a05Wa.wav датасета EmoDB (мужчина, гнев) двумя способами:
   1. Пофреймовая запись в файл 23х низкоуровневых дескрипторов
   2. Вычисление всех параметров на плавающих отрезках длиной 0.2сек с шириной шага 0.16сек
2. Визуализация полученных данных: построение heatmap для каждой матрицы и графиков зависимости от времени каждого признака

В результате ожидаем получить картину, которая поможет делать вывод о целесообразности или нецелесообразности применения сверточных нейронных сетей.