Федеральное государственное образовательное бюджетное учреждение

высшего профессионального образования

**Финансовый университет при Правительстве Российской Федерации**

**(Финансовый университет)**

**Департамент анализа данных и машинного обучения**

Пояснительная записка к курсовой работе

(по дисциплине «Технологии анализа данных и машинного обучения»)

на тему:

**«Машинное обучение в задачах идентификации личности по голосу»**

**Выполнил:**

Студент группы ПИ19-3

Капралов Никита Евгеньевич

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Подпись)

**Проверил:**

Диденко Александр Сергеевич

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Подпись)

**Москва**

**2022**

Оглавление

[Введение 3](#_Toc102620641)

[1. Теоретическая часть 4](#_Toc102620642)

[1.1. Что такое машинное обучение 4](#_Toc102620643)

[1.2. Как работает машинное обучение 4](#_Toc102620644)

[1.2.1 Обучение с учителем 4](#_Toc102620645)

[1.2.2 Обучение без учителя 5](#_Toc102620646)

[1.3. Как охарактеризовать звук 6](#_Toc102620647)

[1.3.1 Эквалайзер 6](#_Toc102620648)

[1.3.2 Хромограмма 7](#_Toc102620649)

[1.3.3 RMS 8](#_Toc102620650)

[1.3.4 Спектральный центроид и спектральная ширина 9](#_Toc102620651)

[1.3.5 Волатильность звука 9](#_Toc102620652)

[1.3.6 MFCC 10](#_Toc102620653)

[2. Практическая часть 12](#_Toc102620654)

[2.1. Используемые технологии 12](#_Toc102620655)

[Librosa 12](#_Toc102620656)

[Matplotlib 12](#_Toc102620657)

[Pandas 12](#_Toc102620658)

[Numpy 13](#_Toc102620659)

[Scikit-Learn 13](#_Toc102620660)

[Keras 13](#_Toc102620661)

[2.2. Определение датасета 13](#_Toc102620662)

[2.3. Определение параметров аудиофайлов 15](#_Toc102620663)

[2.4. Корреляционная матрица 17](#_Toc102620664)

[2.5 Разделение данных 18](#_Toc102620665)

[2.6 Настройка модели 18](#_Toc102620666)

[2.7 Проверка работы модели 20](#_Toc102620667)

[Заключение 22](#_Toc102620668)

[Источники информации 23](#_Toc102620669)

# Введение

Процесс автоматизации и компьютеризации является одним из наиболее значимых явлений мирового прогресса на сегодняшний день. Благодаря нему человечество преодолело необъятное количество трудностей и увеличило качество жизни среднестатистического гражданина в разы. Благодаря автоматизации процессов мы можем делегировать однотипные и рутинные задачи машине, которая справится с поставленной задачей так же хорошо, как человек, а иногда лучше человека. Научные деятели всего мира модернизируют существующие модели автоматизации процессов и создают новые, открывая ранее неявные горизонты и преодолевая неразрешимые трудности.

Задача идентификации личности по голосу является одной из наиболее важных задач машинного обучения на текущий момент. Уже сейчас модели распознавания личности по голосу частично используются во многих сферах в качестве дополнительной верификации личности. Например, в финансовых учреждениях.

Системы идентификации личности по голосу определяют индивидуальные особенности звучания тембра голоса определенного человека. Каждый человеческий голос индивидуален и неповторим, словно отпечаток пальцев.

Цель работы: разработка алгоритма для идентификации личности по голосу.

# Теоретическая часть

## 1.1. Что такое машинное обучение

**Машинное обучение** — это применение искусственного интеллекта, которое помогает моделям машинного обучения учиться и прогрессировать на основе накопленного в процессе тренировки опыта.

**Обучение** — самый важный этап в процессе работы модели. От него зависит точность сгенерированных моделью результатов.

Машинное обучение подразумевает совершенствование модели без вмешательства в процесс обучения человека. Модель совершенствуется без явного программирования.

Система обучается на входных данных, подмечая в них закономерности и корреляцию, благодаря чему и увеличивается точность работы модели. Также важно определить количество входных данных и их параметры для более точного обучения модели.

## 1.2. Как работает машинное обучение

Модель возможно обучить двумя способами. С учителем или без учителя.

### 1.2.1 Обучение с учителем

О**бучение с учителем** — это процесс, в котором система обучается на известных входных и выходных данных, благодаря чему модель легче определяет предсказываемый результат.

В процессе обучения модели с учителем система начинает предсказывать заведомо известные верные данные на основе закономерностей входных параметров. Можно сказать, что модель подстраивается под известные ей выходные данные.

Как правило, данный тип обучения используется в моделях, перед которыми стоит задача что-либо спрогнозировать (например, как в нашем случае, личность, которой принадлежит определенный голос).

### 1.2.2 Обучение без учителя

**Обучение без учителя** — это процесс, который обучает систему находить скрытые паттерны или внутренние структуры во входных данных, не обращая внимания на выходные данные.

В процессе обучения модели без учителя система определяет закономерности и классифицирует данные, не имея точного примера верных выходных данных (такого примера может не существовать). Проще говоря, модели не предоставляют верных ответов в качестве показателя верных выходных данных.

Подобного рода обучение принято использовать в задачах кластеризации.

**Кластеризация** — это процесс автоматического объединения различных точек данных, обладающих сходными характеристиками, и отнесения их к «кластерам».

К примеру, тип рекламных рекомендаций для интернет-пользователей определяется на основе работы модели, обученной без учителя. Нельзя определить со стопроцентной точностью какое именно рекламное объявление больше всего заинтересует пользователя, в связи с этим модель обучается без учителя.

## 1.3. Как охарактеризовать звук

**Звук** — это механическое колебание материи, вызванное вибрацией или колебанием источника. Мы определяем звук в момент того, как он достигает нашего уха и воздействует на барабанную перепонку.  
 Человеческое ухо воспринимает наиболее важный диапазон частот колебания волны: от 20 Герц до 20000 Герц.

Человеческая речь состоит из множества звуковых волн разной частоты. Именно по этой причине мы разделяем голоса различных людей. Каждый голос индивидуален по многим параметрам. Для наглядности можно определить основные частотные волны голоса человека, определяющие восприятие голоса.

**200 Герц** — Определяет глубину голоса.

**600 Герц** — основная частота голоса, определяющая легкость восприятия речи.

**1500 Герц** — определяет чёткость восприятия речи.

**8000 Герц** — частота сибилянтов. Воспроизводит все шипящие звуки.

### 1.3.1 Эквалайзер

Мы определились с тем, что частотность каждого голоса уникальна. Охарактеризовать частотный диапазон голоса можно с помощью инструмента под названием «**Визуальный эквалайзер**». Стоит уточнить, что точность воспроизводимых на устройстве частот зависит от качества микрофона. Даже профессиональные и дорогие микрофоны вносят в процессе записи некие неточности, меняя при этом воспроизводимую частотную волну при воспроизведении записанного материала.

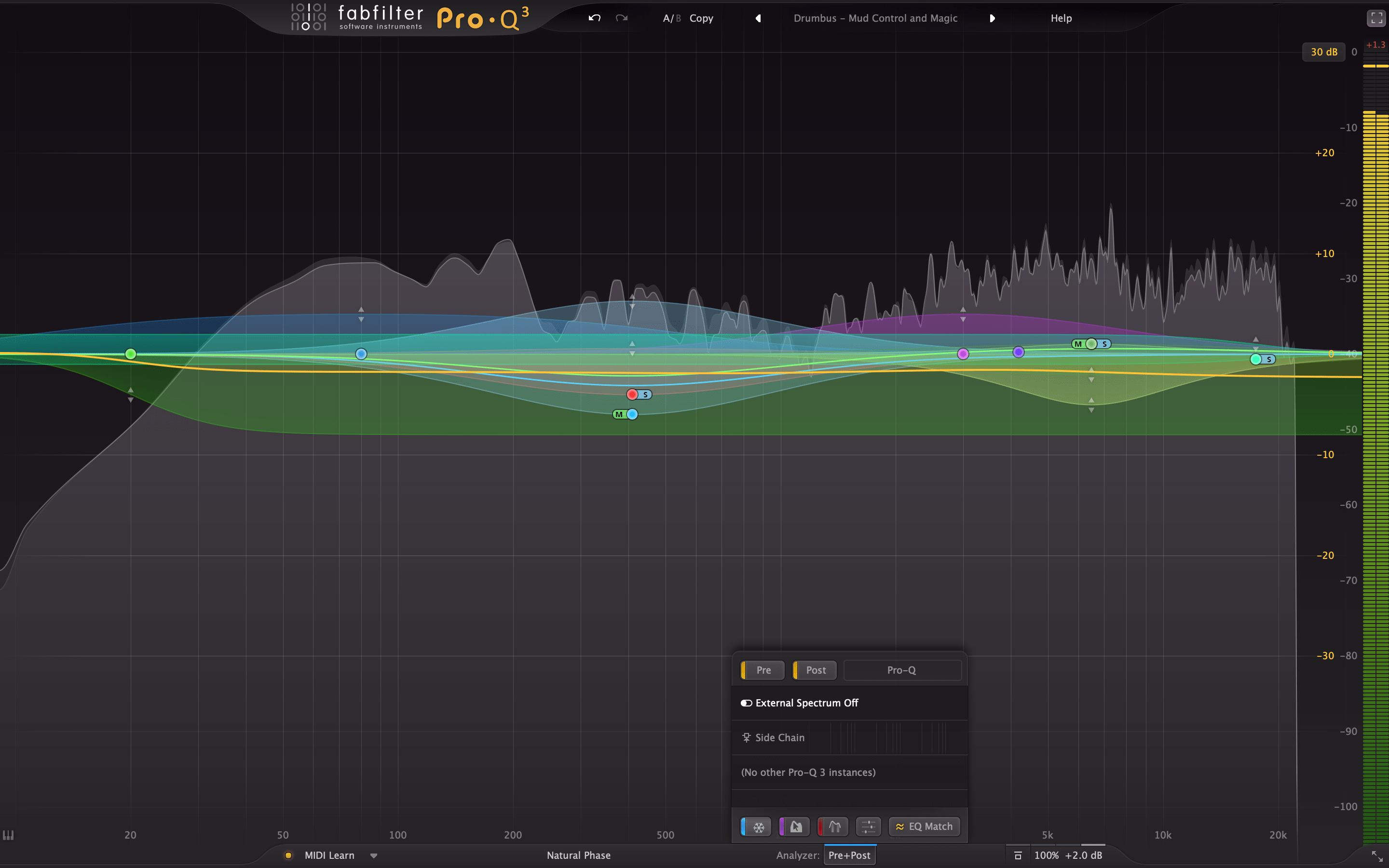


Рис. 1.3.1.1 — эквалайзер FabFilter Pro-Q3

### 1.3.2 Хромограмма

Для более точного определения частотного диапазона звука в реализованной модели используется среднее арифметическое значение показателей хромограммы временного отрезка речи.

**Хромограмма** — вид спектрограммы, наглядно показывающий совокупность амплитуды и мощности частот на определенном временном отрезке воспроизводимого звука. На оси X указывается время, на оси Y указывается частота, а на оси Z указывается амплитуда (хромограмма двумерна, в связи с этим ось Z обозначается определенным цветом, указанном на шкале амплитуды. Амплитуда определяется в Децибелах).

Изображение выглядит как текст, занавеска

Автоматически созданное описание

Рис. 1.3.2.1 — пример хромограммы

### 1.3.3 RMS

Мы определились с видом анализа частотных диапазонов человеческой речи и их визуализации. Теперь стоит уделить внимание громкости человеческой речи. Человек воспроизводит речь тише или громче, как правило, основываясь на внешних факторах, но зависимость точности определения личности спикера от громкости речи присутствует (в последствии будем периодически использовать определение «сигнал» вместо слова «речь»). Согласитесь, что определить личность говорящего, если он шепчет или кричит, гораздо сложнее, нежели он будет говорить на привычной ему громкости.

Использовать для определения громкости сигнала Децибелы не получится, так как этот параметр неточен и непрактичен в решении поставленной задачи.

**Децибел** — дольная единица, равная одной десятой единицы Бел. В основе единицы лежит десятичный логарифм.

Наша модель будет принимать среднеквадратичное значение громкости звукового сигнала: RMS.

**RMS** — это среднеквадратическое значение громкостей всех семплов дорожки.

**Семпл** — малая часть оцифрованного звукового фрагмента.

Использование параметра RMS позволяет определять не пиковое значение громкости сигнала, а среднеквадратичное, что позволяет определить громкость не в определённой точке звукового сигнала, а всего промежутка волны.

Изображение выглядит как текст

Автоматически созданное описание

Рис. 1.3.3.1 — формулы вычисления RMS

a – отдельный семпл

n – сумма семплов

### 1.3.4 Спектральный центроид и спектральная ширина

Человеческая речь имеет определенную основную частоту, которая передаёт наибольшее количество информации о характеристике голоса. Для нахождения данной частоты найдём спектральный центроид сигнала.

Спектральный центроид указывает на какой частоте сосредоточена энергия спектра, или, другими словами, опознаёт, где расположен «центр масс» сигнала.

На основе спектрального центроида также вычисляется дополнительный параметр под названием «спектральная ширина».

**Спектральная ширина** — объём частот, входящих в спектральный центроид.

### 1.3.5 Волатильность звука

Волатильность сигнала напрямую связана с показателями отношений значений эквалайзера и RMS. Сигнал спикера с высокочастотным голосом гораздо менее устойчив и постоянен чем сигнал спикера с более низкочастотным голосом. Как описывалось выше, звуковая волна — это совокупность колебаний какой-либо материи. В нашем случае: воздуха. Если передаваемая волна относится к низкочастотному диапазону, то на графике, изображающим движение волны, данная волна пересечёт ось абсцисс меньшее количество раз, нежели высокочастотная волна, так как ширина и амплитуда её волны будут больше, а количество пересечений с осью абсцисс вследствие этого меньше.

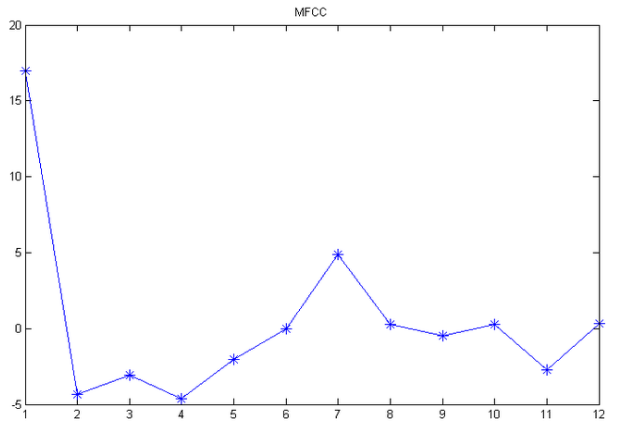
Под волатильностью голоса определяется показатель пересечений совокупности звуковых волн оси Ox на графике волны. Разные голоса имеют разную волатильность.

### 1.3.6 MFCC

Порог слышимости частот человеческого уха неидеален. Если провести эксперимент, при котором подопытному воспроизвести различные звуковые частоты одинаковой абсолютной громкости, то испытуемый выскажет предположение, что низкочастотные звуки имеют меньшую громкость, чем высокочастотные.

Это связанно с тем, что мы воспринимаем низкочастотный звуковой сигнал гораздо хуже, чем высокочастотный. Именно поэтому сирена, срабатывающая в момент экстренной ситуации, кажется нам такой громкой. Она воспроизводит звук в наиболее восприимчивом для человека частотном диапазоне.

Мел-кепстральные коэффициенты позволяют охарактеризовать волну с точки зрения восприятия её человеком. Применив дискретное косинусное преобразование относительности звукового сигнала к показателям восприимчивости звуковых волн, мы получим Мел-кепстральные коэффициенты (MFCC).

Рис. 1.3.6.1 — пример показателей MFCC

# Практическая часть

## Используемые технологии

Python является самым популярным выбором среди множества языков программирования в сфере машинного обучении. В первую очередь в связи с необъятным количеством сторонних библиотек, позволяющих реализовывать необходимые алгоритмы с меньшим количеством написанного кода и ускорением быстроты работы модели. Также сторонние библиотеки легко интегрировать в проект.

Благодаря написанным библиотекам процесс разработки становится быстрее и комфортнее.

### Librosa

**Librosa** — библиотека, ориентированная на анализ и обработку аудиофайлов. С помощью её функционала мы можем извлечь аудиофайл из директории, обработать его и выявить необходимые параметры.

### Matplotlib

**Matplotlib** позволяет визуализировать данные в двумерной и трёхмерной графике. В процессе работы с аудиофайлами необходимо выводить информацию о звуковой волне графически, что реализуемо благодаря Matplotlib.

### Pandas

**Pandas** — библиотека для анализа и структуризации файлов. Крайне удобно параметрически визуализировать и редактировать параметры датасета с помощью Pandas.DataFrame, в связи с чем использование данной технологии в процессе анализа данных является отличным решением.

### Numpy

**Numpy** — библиотека с открытым исходным кодом, предоставляющая возможность удобного варианта работы с многомерными массивами и большим количеством математических параметров. Модель обучения отлично принимает в качестве входных данных объект Numpy.ndarray.

### Scikit-Learn

**Scikit-Learn** — библиотека для предиктивного анализа данных. Используется для предварительной подготовки модели.

### Keras

**Keras** — основная библиотека для обучения модели. Keras позволяет определить и настроить модель, а входящий в Keras пакет модулей Tenserflow позволяет обучить модель в облаке.

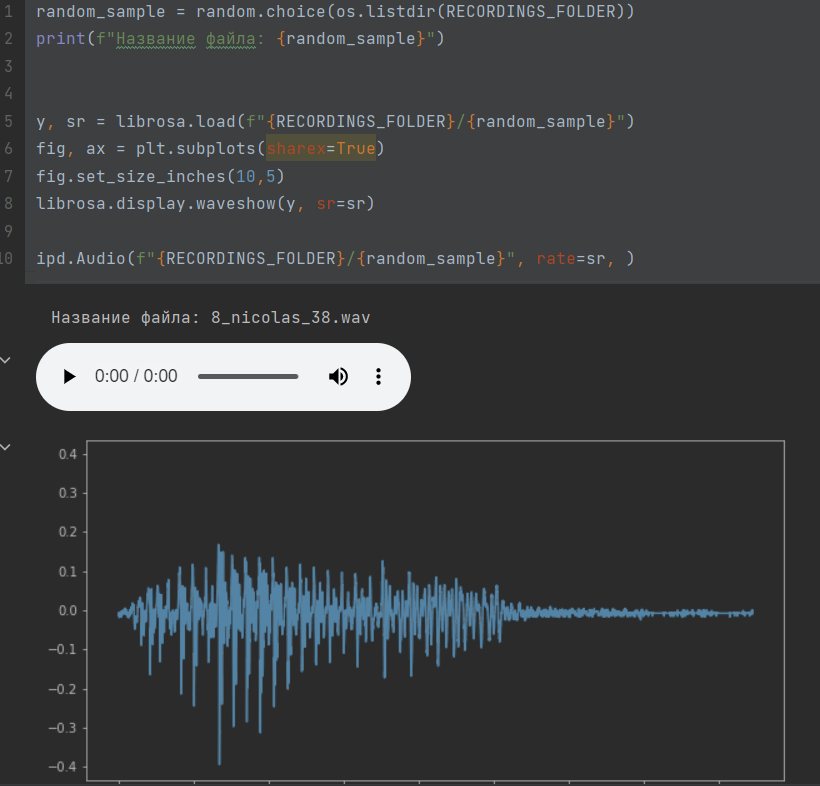
## Определение датасета

Исходный датасет подготовлен техническим университетом Торонто и состоит из 3000 аудиофайлов, содержащих различное произношение цифр от нуля до девяти от шести англоязычных спикеров.

**Спикеры:**

* George
* Jackson
* Lucas
* Nicolas
* Theo
* Yweweler

В качестве первичной проверки датасета выведем информацию о случайном файле из основной директории и прослушаем его.

Рис. 2.2.1 — вывод волны аудиофайла

## Определение параметров аудиофайлов

Для корректной работы модели необходимо передать информацию о параметрах аудиофайлов. Параметры, передаваемые в модель, те же, что указаны в теоретической части пояснительной записки. Также представим данные в виде объекта Pandas.Dataframe. Проверим датафрейм на наличие NULL ячеек.

**Параметры:**

* Среднее хромограммы
* RMS
* Спектральный центроид
* Ширина спектрального центроида
* Волатильность
* MFCС

Изображение выглядит как стол

Автоматически созданное описание

Рис. 2.3.1 — параметры датасета

## Корреляционная матрица

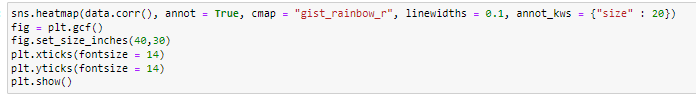
Построим корреляционную матрицу параметров. Корреляционная матрица позволяет найти мультиколлинеарность параметров и увидеть её наглядно на графической визуализации.

Рис. 2.4.1 — вывод корреляционной матрицы

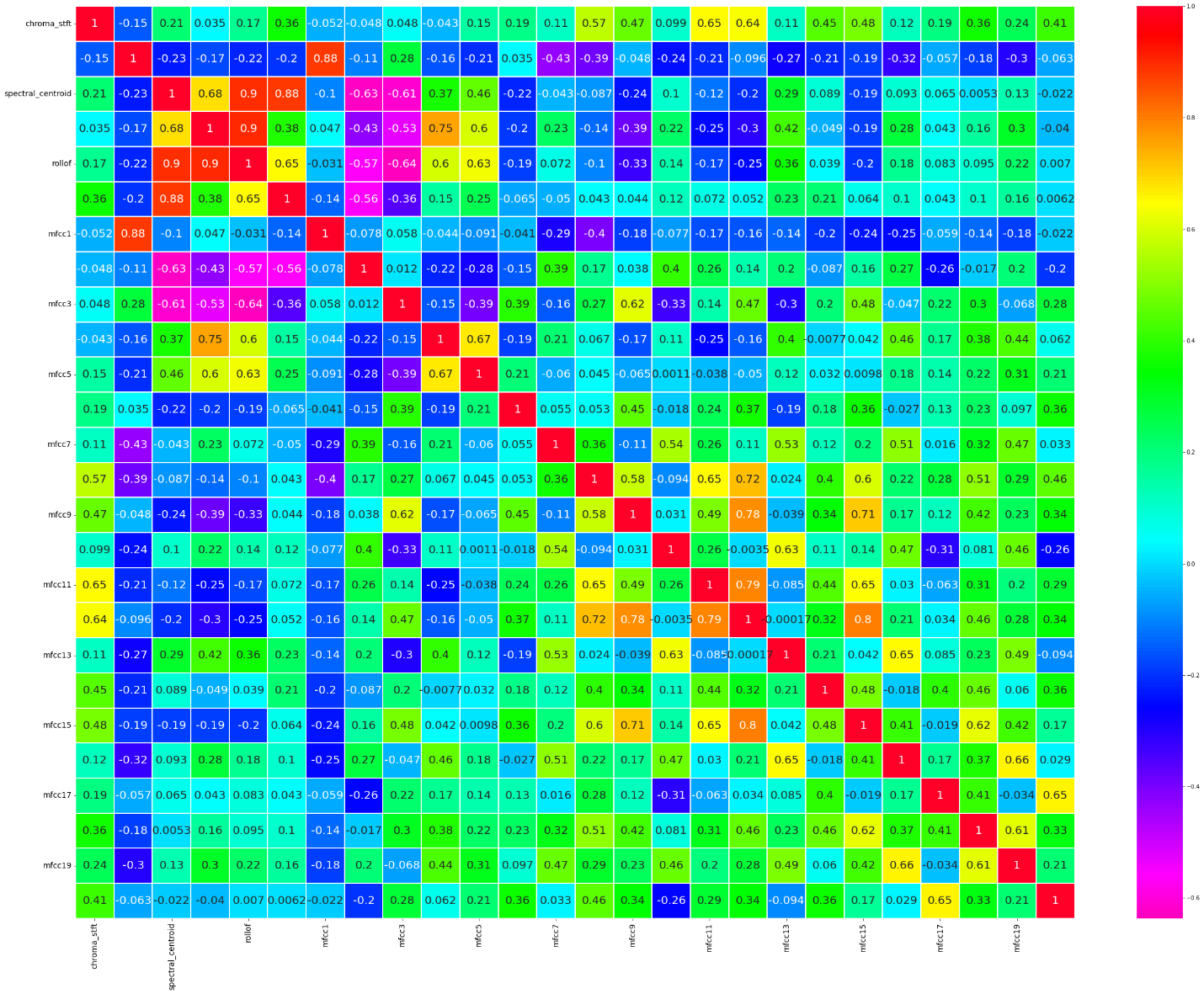


Рис. 2.4.2 — корреляционная матрица

## Разделение данных

После определения основных признаков модели полученные данные необходимо представить в виде Numpy.ndarray и разделить на 3 выборки: тренировочную, валидационную и тестировочную. Для проверки размерности выборок выведем её.

Изображение выглядит как текст

Автоматически созданное описание

Рис. 2.5.1 — разделение датасета

Перед тем как начать настройку модели необходимо полученные данные стандартизировать. Для этого используем метод пакета sklearn.preprocessing под названием StandartScaler.

Изображение выглядит как текст

Автоматически созданное описание

Рис. 2.5.2 — стандартизация датасета

## Настройка модели

Перейдём к настройке модели. В качестве линейного стека слоёв будем использовать три слоя бинарной классификации и один слой мульти классовой классификации. Также определим долю исключений для предотвращения оверфиттинга (переобучения). После проведения вышеуказанных манипуляций скомпилируем модель.

Рис. 2.6.1 — создание модели Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание

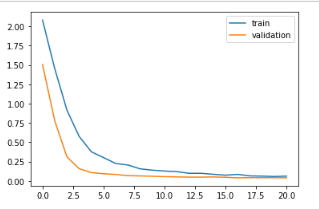
Рис. 2.6.2 — компиляция модели

Обучаем модель. Определяем переменную stopper для остановки обучения при отсутствии прогресса параметра «validation loss». 100 Эпох. Ограничиваем пакет входящих данных в модели в размере 100 единиц.

Изображение выглядит как текст

Автоматически созданное описание

Рис. 2.6.3 — обучение модели

Рис. 2.6.4 — результат обучения модели Изображение выглядит как стол

Автоматически созданное описание

Рис. 2.6.5 — эпохи обучения

## 2.7 Проверка работы модели

Проверим работу модели на случайном датасете.

Изображение выглядит как текст

Автоматически созданное описание

Рис. 2.7.1 — вывод выходных данных модели

Изображение выглядит как стол

Автоматически созданное описание

Рис. 2.7.2 — выходные данные модели

Выведем отчёт о работе модели.

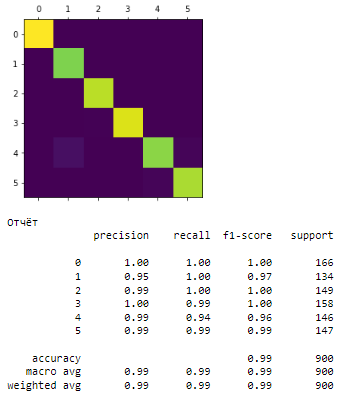


Рис. 2.7.3 — отчёт

# Заключение

Идентификация личности по голосу позволит облегчить процесс подтверждения личности во многих сферах в жизни. Благодаря данной технологии появится возможность идентифицировать человека дистанционно. Например, если человек физически не имеет возможности передвигаться или находиться в недоступной от места нахождения подтверждения личности точки.

Также технологии распознавания личности по голосу могут использоваться в медиапространстве. К примеру, слушая подкаст, появится возможность видеть на экране имя спикера.

Объединение искусственного интеллекта и распознавания голоса вместе оказывается выигрышной стратегией для анализа большого количества аудиофайлов, в то время как программное обеспечение для распознавания одновременно собирает информацию и определяет личность спикера. На практике это означает, что инструменты искусственного интеллекта могут определять личность независимо от человека, обеспечивая оптимизированное управление и экономя время.

# Источники информации

1. Дэви Силен, Арно Мейсман, Мохамед Али. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.
2. Ш.Бастиан. Крупномасштабное машинное обучение вместе с Python. – М.: ДМК Пресс, 2017. – 358 с.
3. Орельен Жерон. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. Концепции, инструменты и техники для создания интеллектуальных систем. – М.: Вильямс, 2018. – 688 с.
4. Себастьян Рашка. Python и машинное обучение. – М.: ДМК Пресс, 2017. – 418 с.
5. Документация Pandas: <https://pandas.pydata.org/docs/>
6. Документация Keras: <https://keras.io/api/>
7. Документация Librosa: https://librosa.org/doc/latest/index.html