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**

# Введение

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

# 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 модели легче определить предсказываемый результат.

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

Подобного рода обучение принято использовать в задачах кластеризации.

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

Тип рекламных рекомендаций для интернет-пользователя определяется на основе работы модели обученной без учителя. Нельзя определить со 100% точностью какое именно рекламное объявление больше всего заинтересует пользователя, поэтому модель обучается без учителя.

## 1.3. Как охарактеризовать звук

Звук – это механическое колебание, 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.

1500 Герц – определяет чёткость восприятия речи

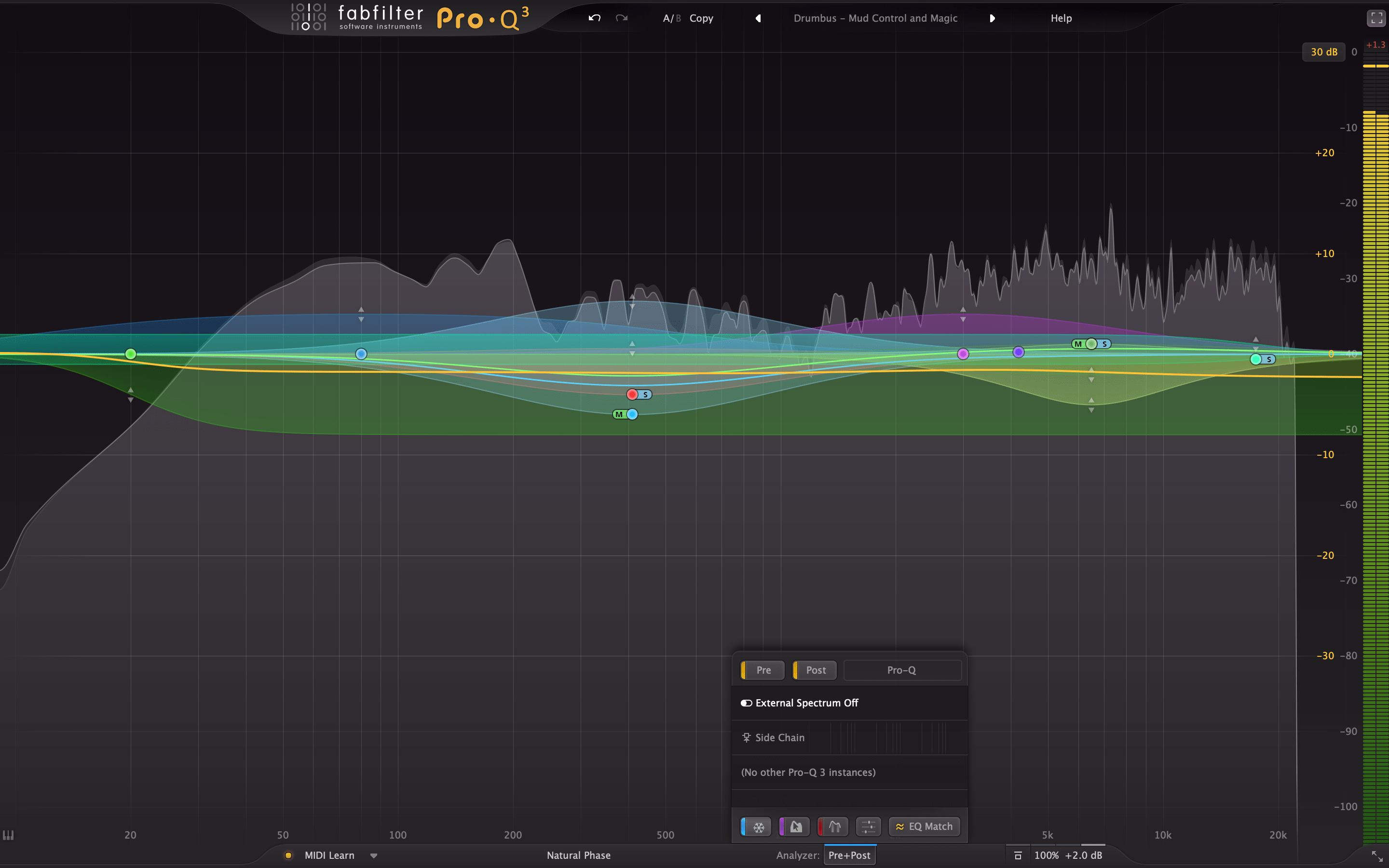
200 Герц – Определяет глубину голоса.

600 Герц – основная частота голоса, определяющая легкость восприятия речи

8000 Герц – частота сибилянтов. Воспроизводит все шипящие звуки

### 1.3.1 Эквалайзер

Мы определились с тем, что частотность каждого голоса уникальна. Определить 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 визуализации.



### 1.3.2 Хромограмма

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

Хромограмма – вид спектограммы, 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 мощности. Амплитуда определяется в Децибелах).

Изображение выглядит как текст, занавеска

Автоматически созданное описание

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 (в последствии будем периодически использовать определение “сигнал” вместо “речь”) присутствует. Согласитесь, что определить личность говорящего, если он шепчет или кричит, гораздо сложнее, нежели он будет говорить на привычной ему громкости

### 1.3.3 RMS

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

Изображение выглядит как текст

Автоматически созданное описание

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.

### 1.3.4 Спектральный центроид

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, указывает, где расположен “центр масс” для звука.

В основе спетрального центроида вычисляется спектральная ширина. Спектральная ширина – обьём частот, входящиг в спектральный центроид.

### 1.3.5 Волатильность звука

Волатильность сигнала напрямую связана с показателями отношением показателя 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 передоваемая волна относится к низкочастотному диопазону, то графике, изображающим движение волны, такая волна пересечёт ось абсцисс меньшее количество раз, нежели высокачастотная волна, так как ширина её волны будет больше, а амплитуда в следствии этого меньше.

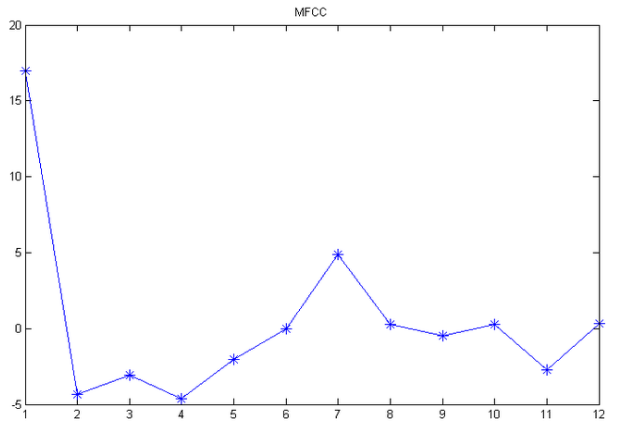
Под волатильностью голоса определяется показатель пересечений совокупности звуковых волн оси Ox на графике волны. Разные голоса имеют разную волатильность.

### 1.3.6 MFCC

Порог слышимости частот человеческого уха неидеален. Если провести эксперимент, при котором подопытному воспроизвести различные звуковые частоты одинаковой абсолютной громкости, то подопытный 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 коэффиценты позволяют охарактеризовать волну с точки зрения восприятия её человеком. Применив дискретное косинусное преобразование к относительности звукового сигнала к показателям восприимчивости звуковых волн мы получим Мел-кепстральные коэффиценты (MFCC).



# Практическая часть

## Используемые технологии

Python является самым популярным выбором среди множества языков программирования в сфере машинного обучения в первую очередь в связи с необъятным количеством сторонних библиотек, которые легко интегрировать в проект.

Благодаря написанным библиотекам процесс разработки становится быстрее и комфортнее.

Librosa

Librosa – библиотека, ореинтированная 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 структурно 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, позваляющая удобный вариант работы с многомерными массивами и большим количеством математических параметров. Модель обучения отлично принимает в качестве входных данных Numpy.ndarray.

Scikit-Leanr

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.

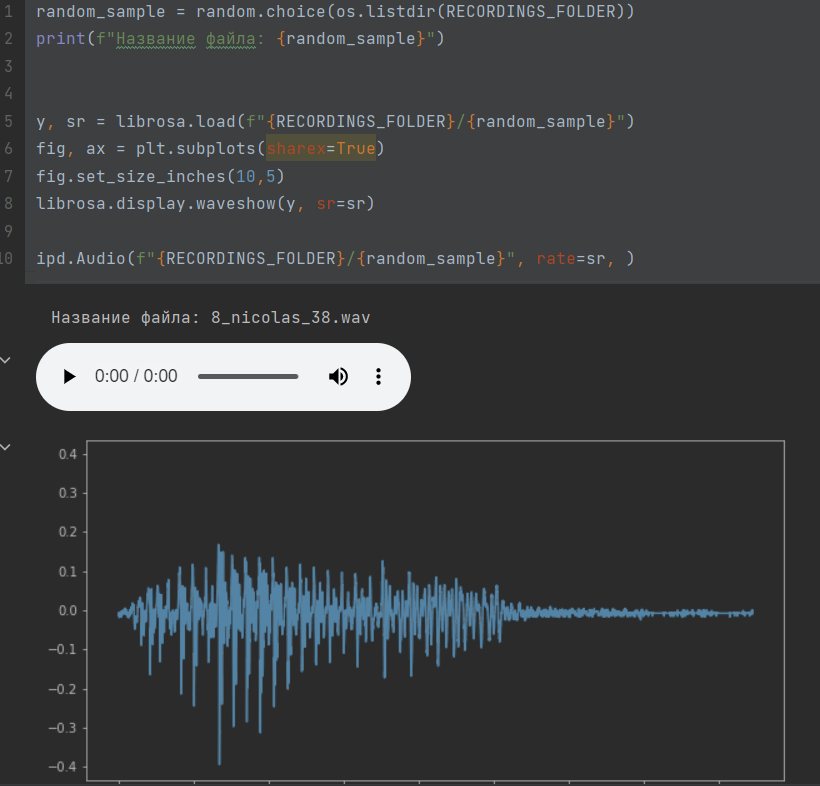
## Определение датасета

Исходный датасет подготовлен техническим университетом Торонто и состоит из 3000 аудиофайлов, содержащих различное произношение цифр от 0 до 9 от 6 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.



## 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

Параметры:

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