УДК 004.8

РАЗРАБОТКА МОДЕЛЕЙ И АЛГОРИТМОВ АУТЕНТИФИКАЦИИ ЧЕЛОВЕКА НА ОСНОВЕ БИОМЕТРИЧЕСКИХ ДАННЫХ (ГОЛОСА)

Териков Е.А., Дзержинский Р.И.

МИРЭА - Российский технологический университет, 119454, Россия, г. Москва, проспект Вернадского, 78, e-mail: terikov232@rambler.ru, 9015111295@mail.ru

Разработана методика создания системы распознавания человека по биометрическим данным голоса. Анализ голоса человека построением мел-спектрограммы, как основы построения нейросети, классифицирующей записи голосов людей. Оценка системы проводится при помощи открытого датасета VoxForge, содержащего размеченные фрагменты речи людей.

Ключевые слова: системы биометрической аутентификации, спектрограмма, мел-спектрограмма, датасет VoxForge, нейросеть, классификатор.

DEVELOPMENT OF MODELS AND ALGORITHMS FOR SPEAKER AUTHENTICATION BASED ON BIOMETRIC DATA (VOICE)

Terikov E.A., Dzerzhinsky R.I

MIREA - Russian Technological University, 119454, Moscow, 78 Vernadskogo Avenue, Russia, e-mail: terikov232@rambler.ru, 9015111295@mail.ru

A method for creating a human recognition system based on biometric voice data has been developed. Analysis of a person's voice by constructing a chalk-spectrogram, as the basis for building a neural network that classifies recordings of people's voices. The system is assessed using the VoxForge open source dataset containing marked fragments of human speech.

Keywords: biometric authentication systems, spectrogram, mel-spectrogram, The VoxForge dataset, neural network, classifier.

Введение

В эпоху цифрового развития и огромных массивов данных, одной из самых важных составляющих информационных систем является сохранность персональных данных о человеке. Для проверки удостоверения личности человека используются системы биометрической аутентификации на основе биометрических данных. Они призваны служить для защиты персональной информации от других людей. Данные системы включают в себя два типа биометрии: статические (данные, которые на основании физиологических особенностей человека не меняются на протяжении всей жизни) и динамические (данные, которые определяются изменением поведения или психологического состояния человека).

Статические виды биометрической аутентификации включают такие показатели, как: отпечатки пальцев, структуру радужной оболочки глаза, сетчатку глаза, геометрию руки, геометрию лица.

Динамические виды биометрической аутентификации представлены голосом и рукописным почерком.

В работе мы рассмотрим методику распознавания человека по голосу, а затем реализуем модель на основе нейронной сети для классификации образцов человеческой речи.

Спектрограмма и мел-спектрограмма

Основной принцип заключается в том, что тембр голоса человека в совокупности с его физиологическими особенностями уникален. Объясняется это множеством факторов, которые влияют на произносимый человеком звук, будь то размер голосовых связок или количество выдыхаемого воздуха. С точки зрения математики голос, как и звук, является волной, у которой есть своя частота и амплитуда. Одним из главных методов анализа голоса является построение мел-спектрограммы и её дальнейший анализ. Под мел-спектрограммой подразумевается особая спектрограмма, учитывающая особенности человеческого слуха. Рассмотрим по порядку данный метод.

Спектр сигнала — это совокупность гармонических колебаний, на которые может быть разложено данное сложное колебание. Спектр представляет входной сигнал в виде набора спектральных линий (частот), определяющих основные компоненты входного сигнала.

Спектрограмма представляет весь регистрируемый акустический спектр сигнала. Звук, как цифровой сигнал, является дискретным и для построения спектрограммы используем **оконное преобразование Фурье**:

$$F(m,w) = \sum_{n=-\infty}^{\infty} f[n]w[n-m]e^{-jwn}$$

где f - функция исходного сигнала,
w - оконная функция,
m - параметр сдвига функции,
n – промежуток сигнала.

На рис. 1 представлен пример спектрограммы фрагмента голоса человека длиной 18 секунд. Ось абсцисс - время (секунды), ось ординат - частота (Гц).

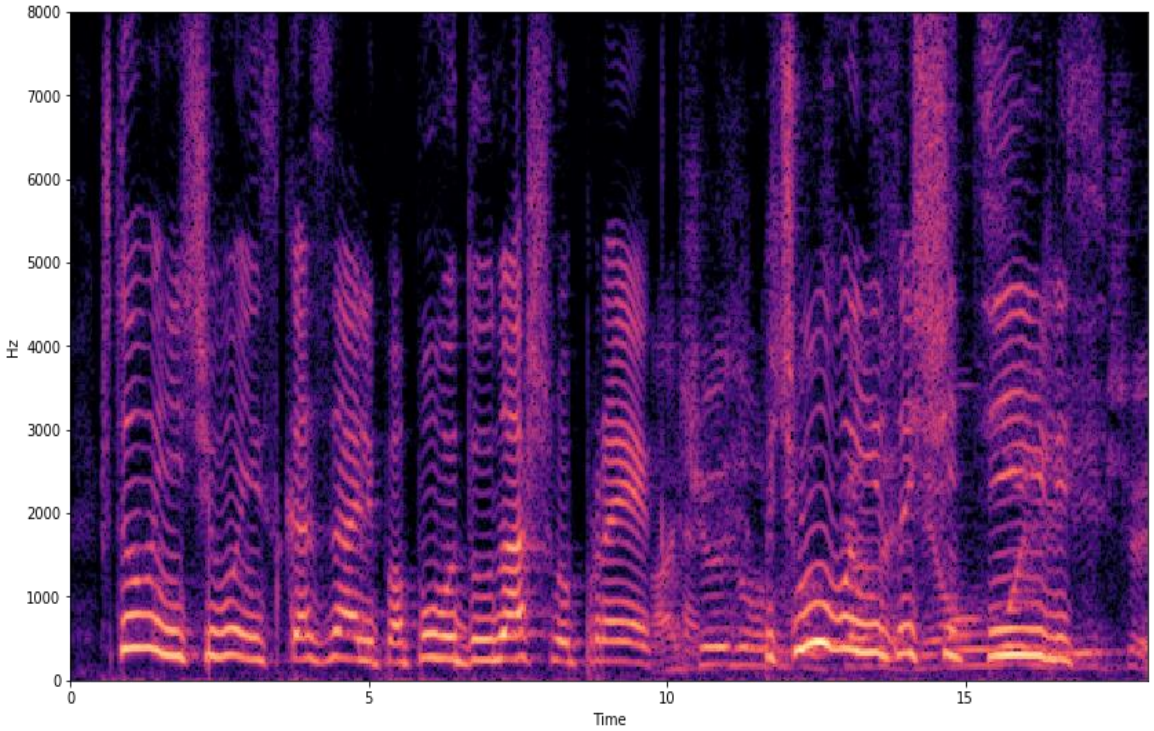

Рис 1. Пример спектрограммы голоса человека

Человеческое ухо приспособлено к восприятию звука на частотах от 16 до 20000 Гц, но при этом есть определённый порог слышимости, за пределами которого человек не способен воспринимать информацию. Чтобы избежать потери полезной информации была введена единица звука, которая могла бы заменить частоту и приблизиться к значениям воспринимаемыми человеком.

Мел – перцепционная единица высоты звука, которая определяется восприятием звукового сигнала органами слуха человека. Зависимость между частотой колебания и высотой звука – нелинейная. Тону с частотой 1 кГц и звуковым давлением

$2*10^{-3}$Па приписывают высоту 1000 мел. Для описания мела существует формула через частоту колебания звука [3]:

$$M = 1125.01048 * ln(1 + \frac{f}{700})$$

где f – частота.

На рис. 2 представлен график зависимости частоты от мела. Ось абсцисс – частота (Гц), ось ординат -мел.

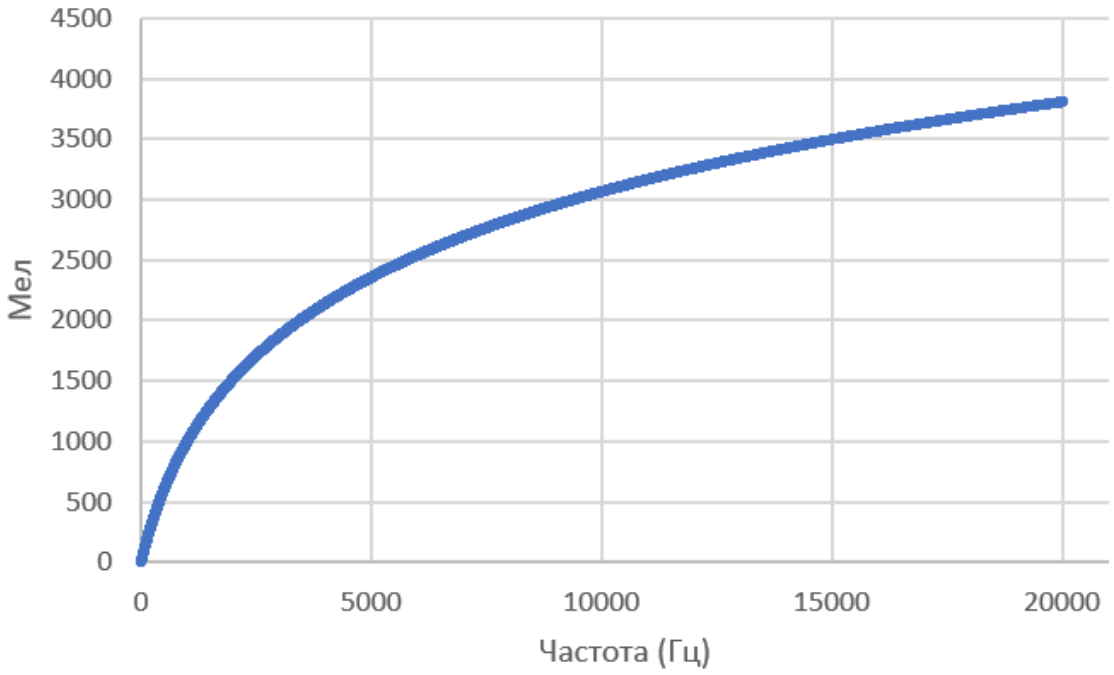

Рис 2. График зависимости мела от частоты звука.

Соответственно для идентификации человека требуется максимально приблизиться к параметрам восприятия человеческого голоса. Для этого к исходной спектрограмме применяются треугольные мел-фильтры. Мел-фильтры — это треугольные функции, которые равномерно, последовательно располагаются на некотором промежутке мел-оси. Это обеспечивает переход от частоты к мелу, а также гарантирует неизменность низких частот и усреднение значений из широкого диапазона для более высоких частот. Соответственно **мел-спектрограмма** — это спектрограмма, в которой частота переведена в меру измерения высоты звука, мел.

При реализации программного комплекса используем архитектуру проекта, представленную на рис. 3.

Рис 3. Архитектура проекта

1. Загрузка файла.

Все входящие файлы, имеющую метку о владельце данного фрагмента голоса, загружаются в формате «.wav» в модуль нормализации и предобработки данных

2. Нормализация данных.

В данном модуле проводится унификация входящие параметров звукового файла, таких как частота дискретизации, длина файла, одноканальная запись звука или двухканальная и др. Нормализация позволяет создать правильную нейросетевую структуру

3. Построение мел-спектрограммы

В данном блоке реализуется построение мел-спектрограммы. На выходе она будет представлять из себя набор векторов

4. Построение модели.

В данном модуле основной задачей будет построение классификатора, задача которого правильно определять каждый следующий поступающий фрагмент голоса

5. Оценка на тестовой выборке

Необходимо результаты работы модели.

Подготовка данных.

Для проверки результата создаётся случайная выборка звуковых файлов из набора данных VoxForge. VoxForge – это открытый набор данных, представляющий собой огромное количество записей фрагментов прочтений книг, новостных зарисовок или диалогов. Так как каждая аудиодорожка помечена, кем она сказана, этот датасет используется для определения личности говорящего.

Для работы с аудио файлами считывается каждый аудиофайл и обязательно делаются метки, какому человеку принадлежит тот или иной фрагмент голоса.

1. Частота дискретизации – 16000 Гц
2. Канал – моно
3. Длительность отрезка – 5 секунд

Теперь данные готовы для их дальнейшего преобразования.

Применив алгоритм получения мел-спектрограммы для каждой записи, получаем набор векторов одной длины, каждый из которых имеет метку о человеке. Модель была реализована на архитектуре, напоминающей адаптацию карты Кохонена [1].

Самоорганизующаяся карта Кохонена – это нейросеть с обучением без учителя, кластеризующая каждый следующий поступающий элемент с корректировкой весов модели. Берётся концепция данной нейросети, и она воссоздаётся по следующему принципу.

Создаётся набор классификаторов, рассчитывающих вероятность принадлежности каждого следующего файла к данному классу, рис. 4.

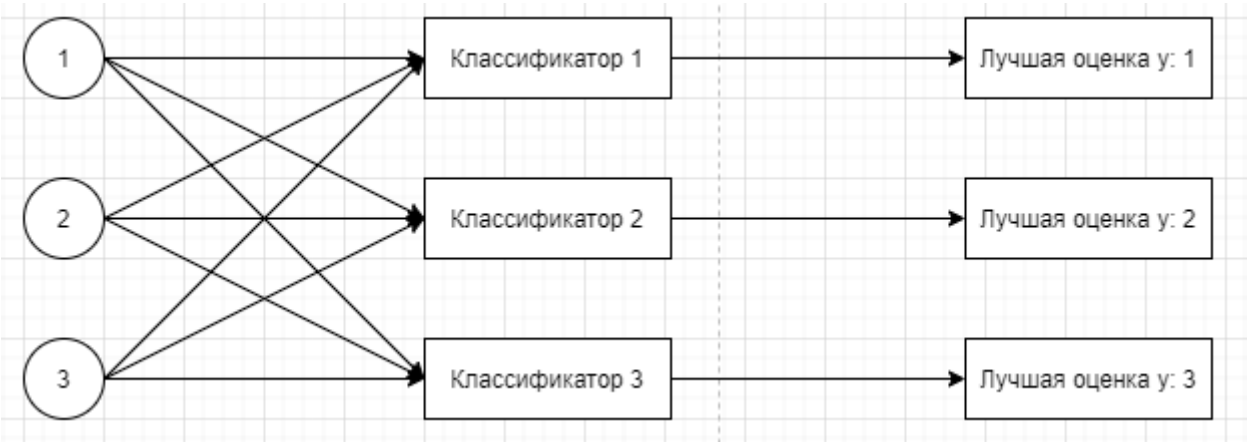

Рис 4. Архитектура классифицирующей нейросети

Для каждого человека, создаётся выборка из 5 аудиофайлов, так, чтобы аудиозаписи могли покрывать наибольшее количество видов произносимых звуков (шипящие, звонкие, свистящие и т.п.). Создаётся количество моделей с одним классом, соответствующее количеству людей, которые на вход будут получать преобразованные аудиофайлы по данному человеку.

На выходе каждая модель даёт свою оценку принадлежности данного экземпляра, и наивысшая оценка является итоговым результатам, и соответственно является распознанным человеком. В роли классификатора выступает модель K – средних [2]. Этот алгоритм был изобретён в 1950 году и его суть в минимизации суммарного квадратичного отклонения точек кластеров от их центров. Именно из-за его архитектуры и рассчитывается евклидово расстояние для оценки вероятности принадлежности данного голоса к тому или другому человеку.

В ходе проведённых экспериментов на датасете VoxForge, была собрана следующая выборка: картотека с 6 записями голосов 60 людей. Тренировочная выборка по 5 записей 60 людей.

Тестовая выборка по 1 записи для 60 людей. На тестовой выборке было правильно идентифицировано 79%.

Заключение

В ходе работы была реализована система определения человека по голосу на основе мел-спектрограмм и нескольких классификаторов К-средних. Данная модель была протестирована на записях VoxForge и отработала с точностью 79%. К положительным моментам можно отнести хорошую точность, и быстродействие алгоритма построения мел-спектрограммы. Минусом данной системы является слишком слабая производительность классификаторов: процесс обучения и самой классификации занимает несколько секунд, что делает его невыгодным для применения в прикладных сервисах. Также хранение большого количества артефактов моделей может негативно повлиять на работу системы. В дальнейшем планируется более глубокое изучение этой темы для оптимизации алгоритмов для работы с большим количеством распознаваемых классов и быстродействием системы.

Список литературы

1. Kohonen T. Self-Organizing Map – 2001.
2. Charu C. A. Chandan K. R., Data Clustering – 2013.
3. Douglas O. Speech communication: human and machine – 1987.
4. Рабинер Л. Р. Цифровая обработка речевых сигналов – 1981.
5. В. Н. Сорокин, В. В. Вьюгин, А. А. Тананыкин, Распознавание личности по голосу: аналитический обзор // Информационные процессы - Том 12, №1 – 2012 – 1-30.

References

1. Kohonen T. Self-Organizing Map – 2001.
2. Charu C. A. Chandan K. R., Data Clustering – 2013.
3. Douglas O. Speech communication: human and machine – 1987.
4. Rabiner L. R. Digital processing of speech signals – 1981.
5. Sorokin V. N., Viugin V. V., Tananukin A. A., Speaker recognition by voice: analytic review // Information processes - Volume 12, №1 – 2012 – P. 1-30.