Санкт-Петербургский государственный университет

Факультет прикладной математики - процессов управления

Кафедра технологии программирования

**Костыренко Виталий Сергеевич**

**Выпускная квалификационная работа бакалавра**

**Метод идентификации человека по голосу**

Направление 01.03.02

Прикладная математика и информатика

Научный руководитель,

старший преподаватель

Стученков А. Б.

Санкт-Петербург

2019

**Оглавление**

Введение.................................................................................................................. 3 Постановка задачи.................................................................................................. 5

Обзор литературы................................................................................................... 6

Глава 1. Обзор существующих методов идентификации диктора….............. 9 1.1 Предварительная обработка данных………………………………............. 9 1.2 Извлечение голосовых характеристик……….……………........................ 12 1.3 Алгоритмы построения голосовой модели…………………………...…….13

Глава 2. Теоретический обзор используемых методов...................................... 18 2.1 Предварительная обработка данных ............................................................. 18 2.2 Вычисление коэффициентов MFCC.............................................................. 20 2.3 Построение голосовой модели с помощью GMM........................................ 22

Глава 3. Программная реализация метода идентификации.............................. 25 3.1 Исходные данные……………………............................................................ 25 3.2 Описание работы программы…..................................................................... 25 3.3 Определение числа компонент гауссовой смеси……................................ 28 3.4 Определение незарегистрированных дикторов ........................................ 29 3.5 Результаты работы программы……………………...................................... 30

Выводы…………………………………………………………….………..…....33

Заключение ............................................................................................................34

Список литературы .............................................................................................. 35

**Введение**

Задача распознавания человека по голосу сформулирована более 40 лет назад, однако исследования этой задачи продолжаются в настоящее время. За последние годы качество распознавания речевой информации значительно повысилось, однако проблема все еще далека от решения. Поэтому исследования существующих алгоритмов и поиск новых решений этой задачи очень актуальны.

Задача идентификации по голосу – процесс определения диктора по записанному образцу его голоса через сравнение с шаблонами из базы данных. То есть имеет место соответствие один к N и необходимо определить, кому именно принадлежит голос.

Задача идентификации диктора разделяется на текстозависимую и текстонезависимую.

Как понятно из названий, текстонезависимая идентификация осуществляет определение диктора по произвольной речи. При такой системе словесное содержание записи не важно, а идентификация происходит только на основе голосовых характеристик говорящего. Такой подход более универсален и лучше применим на практике. Например, в некоторых банках такие системы используются в call-центрах для идентификации клиентов. Это позволяет не тратить время на произнесение кодовых слов и позволяет сохранить естественный ход разговора. Также проверку личности клиента можно запускать в любой момент разговора.

Текстозависимые алгоритмы используются из-за их высокой точности, но они не всегда удобны.

Идентификация диктора по голосу может применяться во многих сферах, требующих подтверждения личности говорящего. Например, как уже говорилось ранее, банки могут использовать эту технологию для идентификации клиентов. В определенный момент стало ясно, что использовать персональные данный, ответы на контрольные вопросы по телефону сильно замедляют обслуживание клиентов и доставляет им неудобства. Для любого человека гораздо проще, чтобы его автоматически идентифицировали, чем вспоминать ответы на вопросы.

Первым банком, внедрившим такую технологию, стал Barclays в 2012 году. Их разработчиком была компания Nuance Communications. Их алгоритм производил идентификацию в течение 30 секунд разговора. С тех пор речевые технологии компании используются в таких организациях, как Royal Bank of Canada, Banco Santander, Tangerine Bank.

Среди банков России и СНГ голосовую биометрию первым начал использовать Тиньков Банк в 2014 году.

Также идентификация по голосу может использоваться в судебной практике, наряду с другими биометрическими данными, такими как ДНК, отпечатки пальцев и т.д.

Анализ звуковых данных в криминалистике имеет много возможностей в применении. Например, можно идентифицировать подозреваемых по записи разговора.

Так же речь может быть собрана и передана более простыми и быстрыми способами, чем, например, ДНК.

Современные методы классификации часто требуют значительного количества образцов в обучающей выборке. Особенностью же задачи идентификации диктора часто является относительно небольшой размер обучающей выборки, что связано с объективными трудностями получения большого объема аудиоинформации, от одного человека. Упомянутое обстоятельство делает невозможным применение средств машинного обучения, чувствительных к малому объему обучающей выборки, например, таких, как нейронные сети. В данной работе делается акцент на разработку методики идентификации, учитывающей эту особенность.

**Постановка задачи**

Цель выпускной квалификационной работы состоит в исследовании имеющихся методов и технологий, позволяющих выполнять идентификацию человека по голосу, реализации и тестировании собственного метода идентификации.

Для достижения цели требуется решить следующие задачи:

1. Исследовать существующие методы решения задачи идентификации человека по голосу, а также выявить их проблемы и ограничения.

2. Разработать эффективный алгоритм идентификации человека по голосу, применимый в ситуации с небольшим объемом обучающей выборки. Выполнить программную реализацию разработанного алгоритма.

3. Провести экспериментальное исследование разработанного алгоритма.

**Обзор литературы**

В данном разделе представлен краткий обзор литературы, используемой при написании данной работы и реализации программной системы.

1. F. Z. Chelali, K. Sadeddine, A. Djeradi Speaker identification system using LPC - Application on Berber language // HDSKD international journal. 2015. Vol. 01, №. 02. С. 29 - 46.

В публикации разрабатывается система идентификации на основе LPC (Linear Predictive Coding) голосовых характеристик. Изучаются преимущества и недостатки этого метода.

1. S. Davis, P. Mermelstein Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. // IEEE transactions on acoustics, speech, and signal processing. 1980. №28, 4. С. 357 - 366.

В данной работе сравнивались несколько параметрических представлений акустического сигнала в отношении эффективности распознавания слов в системе непрерывного распознавания речи, ориентированной на слоги.

1. S. Furui Cepstral analysis technique for automatic speaker verification // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1981. №29, 2. С. 254–272.

В этой статье описываются новые методы автоматической проверки говорящего с использованием телефонной речи. В частности, предлагается новый метод DTW (Dynamic Time warping) с использованием техники динамического программирования.

1. D. Ramage Hidden Markov Models Fundamentals // CS229 Section Notes. 2007.

Эти заметки дают подробное математическое введение в концепцию Марковских моделей - формализм для рассуждений о состояниях во времени и скрытых Марковских моделях, где мы хотим восстановить ряд состояний из серии наблюдений.

1. A. Solomonoff, C. Quillen, W. Campbell Channel compensation for svm speaker recognition // ODYSSEY04 -- The Speaker and Language Recognition Workshop Toledo. 2004.

В этой статье исследуются методы, специфичные для структуры SVM (Support Vector Machine), для получения полностью нелинейных компенсаций канала.

1. Иванов И. И. Анализ метода мел-частотных кепстральных коэффициентов применительно к процедуре голосовой аутентификации // Актуальные проблемы гуманитарных и естественных наук. 2015. №10-1. С. 106-114.

В данной статье изучается применение MFCC (Mel-Frequency Cepstral Coefficients) коэффициентов к голосовой аутентификации. Сравнивается качество аутентификации по сравнению с другими коэффициентами с применением моделей HMM (Hidden Markov Model) и SVM.

1. G. Saha, S. Chakroborty, S. Senapati A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications. 2005.

В данной статье предлагается эффективный метод удаления шумов и беззвучных участков в звуковом сигнале.

1. Jeff A. Bilmes A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models // International Computer Science Institute Berkeley CA. 1998. С. 7 - 13.

Здесь описывается проблема оценки параметра максимального правдоподобия и то, как алгоритм максимума ожиданий EM (Expectation Maximization) может использоваться для ее решения.

1. D. A Reynolds, T. F. Quatieri, R. B. Dunn Speaker Verification using adapted Gaussian mixture models // Digital Signal Processing. 2000. Vol. 10. С. 19 - 41.

В этой статье авторы подробно описывают принцип универсальной фоновой модели и ее адаптации для описания моделей дикторов.

**Глава 1. Обзор существующих методов идентификации диктора**

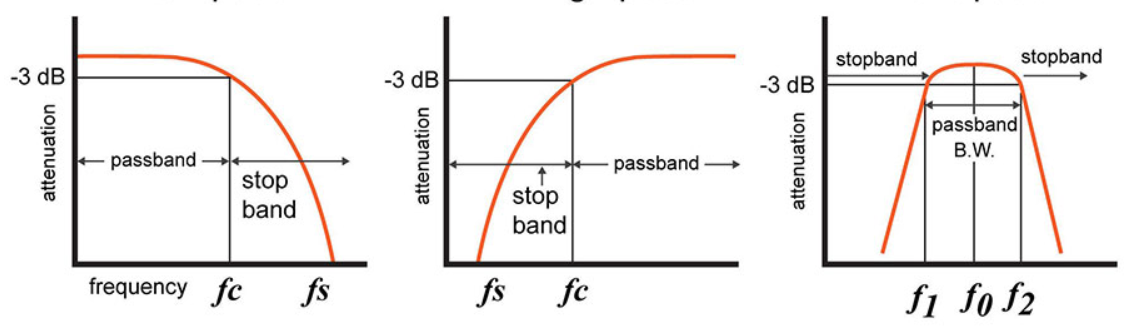
Процесс идентификации диктора можно разделить на три основных этапа: этап предварительной обработки данных (удаление шумов и пауз из записи), извлечение голосовых характеристик и построение голосовой модели диктора. Определение диктора во всех рассматриваемых методах состоит в поиске модели, наиболее близкой к речевому фрагменту диктора, на основе соответствующих модели критериев. В течение последних лет техники этих этапов существенно изменялись, однако наиболее эффективные методы до сих пор используются как фундамент для многих систем обработки речи.

* 1. **Предварительная обработка данных**

Вместе с полезным сигналом в микрофон часто попадают различные шумы, которые отрицательно сказываются на качестве идентификации, поэтому их необходимо удалить.

Классическим методом в устранении шумов является применение частотных фильтров (рис 1.). Основными частотными фильтрами являются:

1. Фильтр высокой частоты (high-pass filter) удаляет из входного сигнала все частоты, значения которых ниже некоторого порога. Частотный диапазон речи начинается с 300 Гц, поэтому все частоты ниже этой можно вырезать из сигнала.
2. Фильтры низкой частоты (low-pass filter) вырезают частоты, которые выше некоторого порога. Для речевого сигнала этим порогом является 4000 Гц.
3. Пропускающий полосовой фильтр (band-pass filter) по сути является комбинацией двух предыдущих, пропуская сигналы, частоты которых содержатся в некоторой допустимой полосе. Аналогично, в качестве нижнего и верхнего порога выбирается соответственно 300 Гц и 4000 Гц.



а) Фильтр низкой частоты б) Фильтр высокой частоты в) Полосовой фильтр

Рис.1

Еще одним важным этапом предварительной обработки является удаление неинформативных участков сигнала. Такими участками являются моменты, когда в сигнале отсутствует голос диктора, например, между словами или в начале и конце записи.

Одним из способов выделения неинформативных участков является метод, основанный на кратковременной энергии сигнала.

Сигнал делится на участки (фреймы) одинаковой длины и для каждого вычисляется его энергия . Затем каждое значение энергии сравнивается с некоторым пороговым значением . Если энергия фрейма выше порогового значения, то фрейм считается активным, что значит, на нем присутствует голос диктора.

Другими способами являются методы, основанные на анализе плотности распределения значений отсчетов паузы. Они базируются на том, что при записи, первый малый по времени (например, 200 мс) отрезок сигнала является паузой. По этому отрезку вычисляется информация о нормальной плотности распределения не озвученного сигнала, которая затем используется для выделения речи.

Один из таких методов подробно рассматривается на странице 16.

Пример применения алгоритмов удаления шумов и выделения речевых участков сигнала показан на рисунке 2.

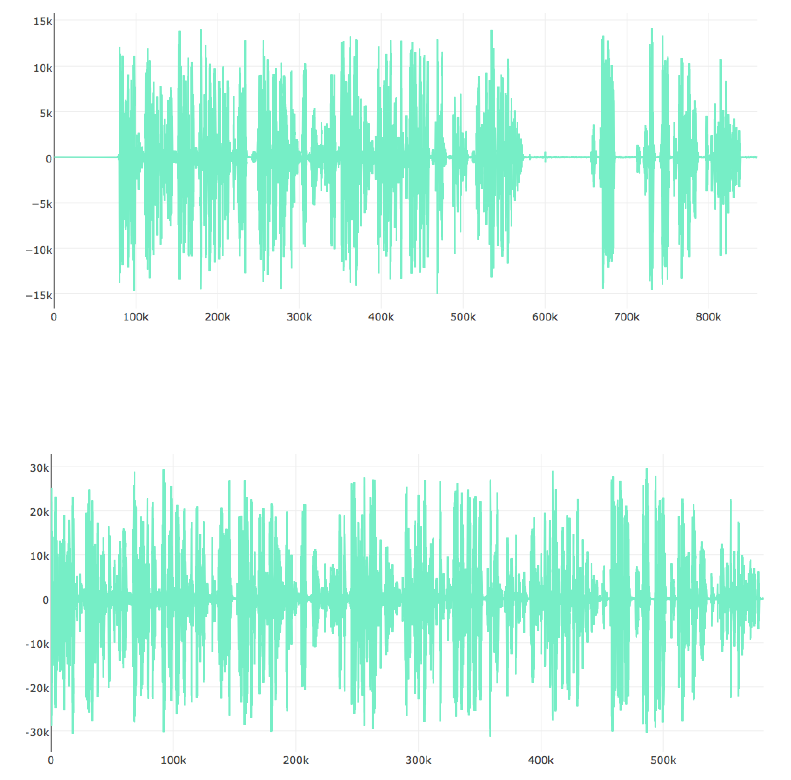


Рис. 2: Сигнал до и после удаления шумов и пауз

* 1. **Извлечение голосовых характеристик**

После предобработки звука мы получаем дискретную форму сигнала. Такие данные не несут в себе достаточно эффективную информацию для дальнейшего анализа, поэтому ее необходимо преобразовать.

В настоящее время существует несколько техник для вычисления голосовых характеристик, которые передают информацию о сигнале:

1. Использование банков фильтров. В данном методе сначала над сигналом производится преобразование Фурье, которое затем делится на фильтр по частотам. На каждой из частот считаются свойства сигналов. Затем, фильтры по частотам складываются в вектор.

Данный метод выдает слишком большой результирующий вектор с плохо выделенными признаками, так как они получены на разных отрезках частот и будут сильно коррелировать, что приведет к потере информации.

Таким образом, данный метод неприменим к задаче идентификации по голосу.

1. Коэффициенты линейного предсказания LPC (Linear Prediction Coding), которые представляют спектральную огибающую голосового сигнала. Этот метод предполагает, что значение сигнала в определенный момент времени является линейной комбинацией предыдущих значений сигнала, что представляется в виде формулы: , где  - коэффициенты линейного предсказания. Ищутся коэффициенты , при которых значение ошибка  будет минимальным.

Запишем, что , и приравняем к нулю, получив систему из p уравнений вида: , где i изменяется от 1 до p.

Далее необходимо решить данную систему линейных уравнений и получить коэффициенты линейного предсказания.

Недостатком данного метода является значительное снижение информативности при наличии шумов [1].

1. Мел-частотные кепстральные коэффициенты MFCC (Mel-Frequency Cepstral Coefficients) являются наиболее популярным выбором при работе со звуком.

Мел – единица измерения частоты звука, основанная на статистической обработке больших данных о субъективном восприятии высоты звуковых тонов.

Кепстр – преобразование, с помощью которого фильтр отделяется от источника звуковой волны.

Особенностью данного подхода является полученного вектора характеристик от длины исходного сигнала и учет в нем разброса индивидуальных особенностей, говорящего [2].

* 1. **Алгоритмы построения голосовой модели**

После извлечения голосовых характеристик каждая запись представлена в виде набора векторов, с помощью которого необходимо обучить алгоритм классификации. Существует множество способов построения голосовой модели диктора:

**Алгоритм динамической трансформации временной шкалы DTW (Dynamic Time Warping)** [3]. Данный алгоритм хорошо подходит для текстозависимых систем распознавания.

В качестве сохраненной модели (модели зарегистрированного диктора) выступает набор векторов признаков речевого сигнала из обучающей выборки .

Пусть  - набор векторов признаков входного сигнала проверяемого диктора.  - матрица выравнивания для двух наборов векторов, в каждой позиции которой содержится значение выравнивания между элементами векторов из С и Q соответственно.  - набор индексов смежных элементов, который определяет соответствие между элементами сопоставляемых векторов. Элементы набора  должны удовлетворять двум условиям:

1. 
2. Если , то , где 

Целью алгоритма DTW является нахождение набора , который удовлетворяет данным выше условиям и при котором было бы минимальным суммарное искажение C относительно Q:



Значение этой формулы будет определять меру близости Q и C. Для его вычисления используется метод динамического программирования, на каждом шаге которого вычисляется значение  по формуле:



При этом , а остальные элементы строки и столбца с индексом 0 приравниваются к ∞.  определяет евклидовое расстояние между элементами i и j. В конце работы алгоритма вычисляется значение .

Определение диктора происходит путем вычисления минимума значения . В итоге получаем номер i диктора, образец голоса которого наиболее близок к образцу голоса входного диктора C.

Преимуществом данного метода является простота реализации, однако он неприменим к текстонезависимой идентификации.

Следующей рассматриваемой моделью является **скрытая Марковская модель HMM (Hidden Markov Model)** – модель для решения задач классификации скрытых параметров по наблюдаемым [4].

Модель HMM определяется следующими параметрами:

*  - множество наблюдений, где  - начальное наблюдение, а  - конечное;
*  - множество скрытых состояний, где - начальное состояние, а  – конечное;
*  - исходное распределение состояний, которое определяет вероятность начать работу в состоянии i;
*  - матрица вероятностей наблюдений, где 
*  - матрица вероятностей переходов между скрытыми состояниями, где 

В рамках задачи идентификации диктора по голосу скрытыми состояниями являются векторы признаков из обучающей выборки, а в качестве наблюдений – векторы признаков тестовой выборки. Сохраняемая модель описывается с помощью матрицы A и B, которые необходимо вычислить с помощью алгоритма Баума-Велша.

Скрытые Марковские модели просты в понимании, имеют большую точность в распознавании, однако применяется она в основном для текстозависимой идентификации.

**Метод опорных векторов SVM (Support Vector Machine)** – бинарный классификатор [5], основная идея которого состоит в разделении пространства признаков двух классов гиперплоскостью, которая задается формулой .

Пусть заданы наборы признаков  и соответствующие им метки двух классов .

В случае, когда данные линейно разделимы, условия нахождения функции  записываются в виде:



Для лучшего разделения классов необходимо найти такую гиперплоскость, при которой достигается максимальное расстояние между ней и множествами обучающих признаков. Расстояние вычисляется как . Таким образом, задача поиска разделяющей гиперплоскости сводится к минимизации  при указанных выше условиях. Эта задача минимизации может быть решена с помощью метода множителей Лагранжа.

В случае линейно-неразделимых множеств необходимо вводить функцию ядра , которая определяет скалярное произведение в новом пространстве. Параметры ядра определяются путем оценки методом кросс-валидации.

В качестве сохраненной модели диктора выступают параметры функции , а также функция ядра. Знак полученного выражения  определяет принадлежность вектора  к одному из двух классов.

В случае многоклассовой классификации обычно используется стратегия «один против остальных», при которой строятся q классификаторов, которые обучаются отличать один класс от остальных, а решение принимается на основе наибольшего значения разделяющей функции ,

У метода опорных векторов высокая точность классификации. Имеется возможность применять разные функции ядра для различных данных. Однако, недостатками являются проблема выбора функции ядра и медленная обучаемость в случае многоклассовой классификации.

Модель гауссовых смесей GMM (Gaussian Mixture Model) является одним из самых популярных способов построения голосовой модели диктора. Исходные данные представляются в виде кластеров, описываемых гауссианами. Система, основанная на этом подходе, не будет чувствительна к временным аспектам речи, и может быть использована для текстонезависимой идентификации диктора. Алгоритм построения модели гауссовых смесей подробно рассмотрен на стр. 22.

**Глава 2. Теоретический обзор используемых методов**

Разрабатываемая технология идентификации на различных этапах опирается на некоторые из алгоритмов и методов, рассмотренных выше. В данной главе детально описывается их применение.

В качестве используемых признаков было решено использовать мел-частотные кепстральные коэффициента MFCC, поскольку их реализация проста, а точность идентификации при их использовании достаточно высока [6]. В качестве голосовой модели была выбрана модель гауссовых смесей GMM из-за ее высокой точности в текстонезависимой идентификации.

* 1. **Предварительная обработка данных**

В качестве предварительной обработки был выбран алгоритм, описанный в статье [6]. В алгоритме предполагается, что фоновый шум имеет гауссовский характер. Основная идея метода в том, что некоторое числе первых миллисекунд записи имеют только лишь шум, так как диктору необходимо время на прочтение. Рассмотрим подробное описание этого алгоритма.

Алгоритм разделен на 5 шагов:

Пусть частота дискретизации равна 8000 Гц. Тогда первым 200 мс записи соответствует первые 1600 дискретных значений сигнала, которые будем называть сэмплами (sample).

Шаг 1: Расчет среднего и стандартного отклонения первых 1600 мс записи. Если µ и σ - среднее значение и стандартное отклонение соответственно, то аналитически можно записать





Шаг 2: На каждом сэмпле вычисляется функция расстояния Махаланобиса, которая сравнивается с числом 3.



Если расстояние больше трех, то сэмпл озвучен.

Шаг 3: Озвученные сэмплы помечаются единицей, а не озвученные нулями. Все сэмплы группируются в последовательные непересекающиеся окна, длиной 10 мс. Теперь весь сигнал представлен нулями и единицами.

Шаг 4: Пусть в каждом окне имеем M нулей и N единиц. Если M ≥ N, то меняем нули на единицу, а единицы на нули. Здесь учитывается, что система речи не может внезапно измениться за такой короткий промежуток.

Шаг 5: Извлекаем только те сэмплы, которые помечена единицами.

Таким образом, метод использует статистические свойства фонового шума, а также физиологические аспекты процесса производства речи. Показано, что он эффективен в вычислительном отношении для приложений реального времени и работает лучше, чем обычные методы для выборок речи, собранных в шумной, а также свободной от шума среде.

* 1. **Вычисление коэффициентов MFCC**

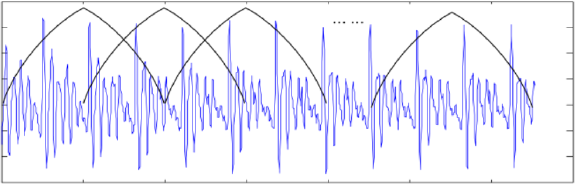
Рассмотрим подробнее извлечение коэффициентов MFCC.

Первым делом дискретный сигнал разбивается на пересекающиеся фреймы, то есть на участки по N миллисекунд с некоторым шагом L (рис. 3а), который обычно равен N/2.

Далее, чтобы сгладить границы фреймов, используют оконную функцию Хэмминга.

Следующий шаг – извлечение спектральной информации с помощью дискретного преобразования Фурье. После этого преобразования получаем набор комплексных чисел, содержащих амплитуду и фазу входного сигнала.

После нужно получить проекцию спектра из частотной шкалы в мел-шкалу. Связь этих величин изображена на рис. 3б и выражается следующей формулой:



а) Пересекающиеся фреймы б) Зависимость мел от частоты

колебаний звука

Рис. 3

Далее формируются треугольные фильтры (рис. 4), которые представляют собой оконные функции.

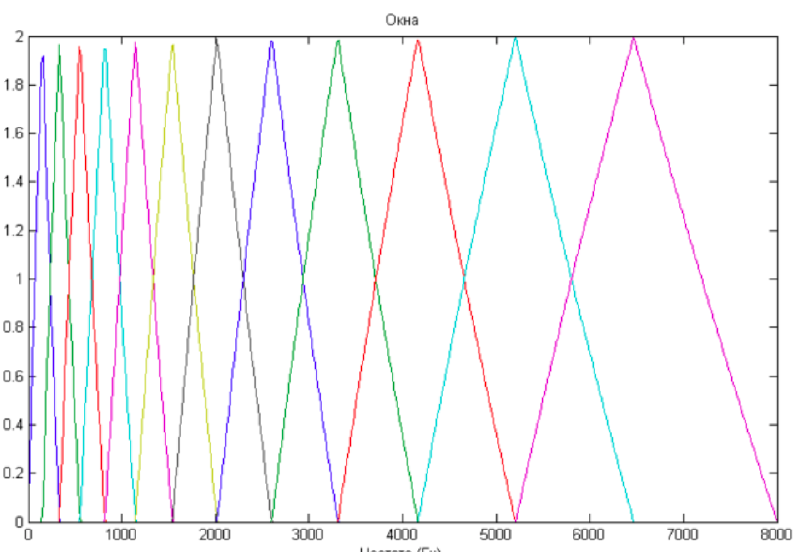


Рис. 4: Мел-фильтры

Перемножив вектора спектра и значения оконной функции, поучим набор спектральных коэффициентов, которые необходимо перевести в кепстральные.

Перевод можно осуществить либо с помощью дискретного преобразования Фурье, либо с помощью дискретного косинусного преобразования. Таким образом, кепстр можно определить следующей формулой:



В этой формуле  - значение сигнала t, ,  - общее количество сигналов.

В качестве результата берутся первые 12 значений кепстра.

После этого для каждого фрейма по формуле ниже вычисляется энергия сигнала в качестве еще одного признака.



Заключительным этапом является вычисление дельта и двойных дельта значений на основе предыдущих признаков, которые также добавляются в результирующий вектор коэффициентов.





Затем, дополнительно производится нормализация для подавления влияния канала и шума на качество распознавания диктора.

* 1. **Построение голосовой модели с помощью GMM**

Как уже говорилось ранее, в модели гауссовых смесей исходные данные представляются в виде кластеров, описываемых гауссианами. В примере на рис. 5, данными являются синие точки, а моделью гауссовых смесей (набором гауссиан) – эллипсы, покрывающие эти данные.

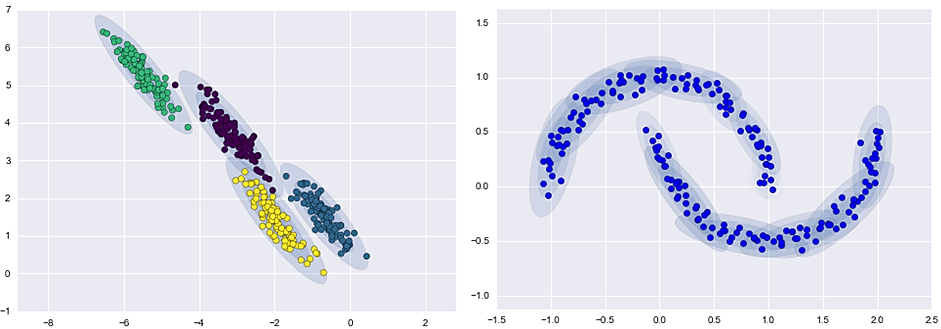


Рис. 5: Модель гауссовых смесей

Модель гауссовых смесей определяется векторами математического ожидания , ковариационной матрицей ∑ и векторов весов  для каждой компоненты смеси.

Таким образом, для построения голосовой модели необходимо оценить эти параметры так, чтобы они наилучшим образом соответствовали распределению векторов признаков говорящего.

Наиболее используемым методом оценки параметров этой модели является метод максимального правдоподобия EM (Expectation- Maximization) [7]. Это итерационный алгоритм, на каждой итерации которого вычисляются следующий значения:







Здесь и  - соответственно новые и старые значения весов, математических ожиданий и ковариационных матриц модели, а в качестве  обозначается модель,  - вектор признаков.

Алгоритм выполняется до схождения параметров или достижения максимального числа итераций. Пусть М – количество компонент, D – размерность вектора признаков. Значения апостериорной вероятности вычисляются по формулам:





Однако, для более быстрой и точной сходимости параметров, необходимо наиболее правильно инициализировать начальные параметры модели. Для этого используют алгоритм кластеризации к-средних (K-means++), после которого уже выполняется EM-алгоритм.

Еще один параметр, который необходимо установить, это количество компонент гауссовой смеси. Этот показатель обычно подбирается экспериментально и в зависимости от имеющихся ресурсов.

В итоге каждый диктор представляется в виде отдельной смеси гауссовых распределений. Идентификация диктора производится путем максимизации апостериорной вероятности:





Используя логарифм и то, что наблюдения независимы между собой, система идентификации диктора в итоге вычисляет:



**Глава 3. Программная реализация метода идентификации**

* 1. **Исходные данные**

В качестве обучающей выборки были собраны записи профессиональных русскоязычных дикторов с различных сайтов предоставления услуг озвучивания. Всего 100 записей в среднем длиной в минуту секунд с частотой дискретизации 16000 Гц. В выборке преобладают мужские голоса, поэтому итоговые результаты могут быть смещены в эту сторону.

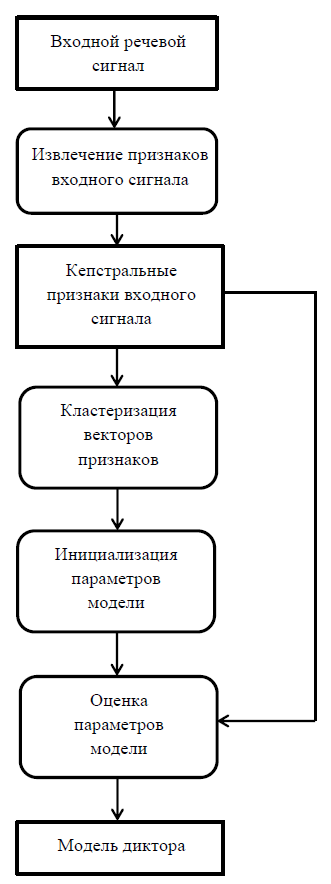
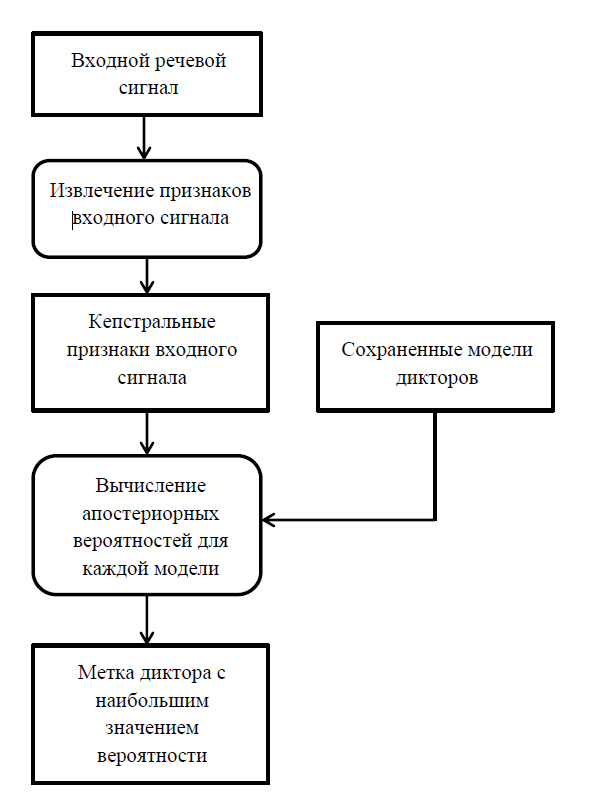
* 1. **Описание работы программы**

Алгоритм идентификации диктора по голосу был реализован на языке Java. Были самостоятельно написаны классы, соответствующие извлечению MFCC коэффициентов и построению голосовой модели диктора. Интерфейс пользователя создан с помощью платформы JavaFX.

Основные классы в программе:

* EndPointDetection – класс, метод doEndPointDetection которого реализует предварительную обработку сигнала, а именно, удаление шумов и не озвученных участков сигнала.
* Классы DCT, FFT, Energy, MFCC и Delta реализуют дискретное косинусное преобразование, быстрое преобразование Фурье, вычисление энергии, мел-частотных кепстральных коэффициентов, дельта и двойных дельта значений для формирования признаков речевого сигнала.
* Классы GMM, GMMTrainer, GMMTrainerParams и KMeansClusteringTrainer и KMeansClusteringTrainerParams реализуют модели и методы обучения смеси гауссовых распределений и K-Means++ кластеризацию, а также методы сохранения моделей в файл формата .txt;
* Последний основной класс в алгоритме Classification, который реализует метод принятия решений при идентификации.

Общую схему работы системы идентификации диктора в режимах обучения и идентификации можно изобразить следующим образом:

а) Обучение модели б) Идентификация диктора

Рис. 6

Пользовательский интерфейс программы выглядит следующим образом:

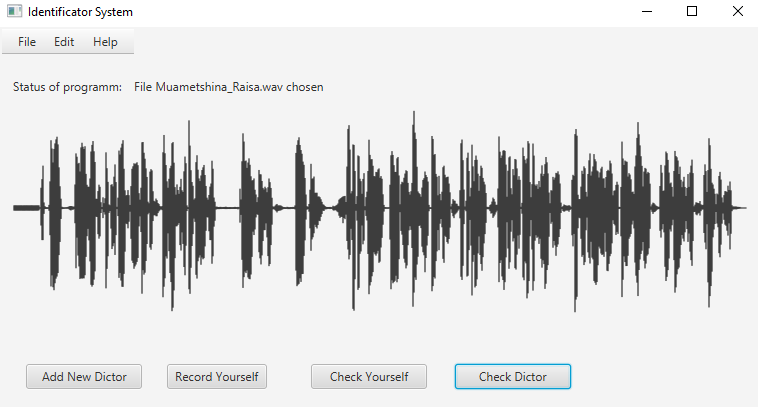


Рис. 7: Пользовательский интерфейс

Он позволяет пользователю зарегистрировать нового диктора как путем выбора отдельного .wav файла с помощью кнопки Add New Dictor, так и записью голоса с помощью микрофона с помощью кнопки Record Yourself. Аналогично для проверки диктора можно использовать файл или микрофон с помощью кнопок Check Dictor и Check Yourself соответственно. Так же после выбора или записи речи диктора, по центру будет выведена звуковая осциллограмма для визуализации записи.

* 1. **Определение числа компонент гауссовой смеси**

Как говорилось ранее, выбор числа компонентов модели является одной из основных проблем при обучении модели гауссовых смесей.

Для определение этого параметра проводилось несколько обучений и тестирований с различными значениями числа компонент. Использовались 25 дикторов из обучающей выборки и 25 тех же дикторов из тестируемой. Результаты при разном числе компонент представлены в следующей таблице:

|  |  |
| --- | --- |
| Число компонент | Точность идентификации, % |
| 1 | 84 |
| 4 | 92 |
| 8 | 92 |
| 16 | 92 |
| 32 | 96 |
| 64 | 96 |
| 128 | 96 |
| 256 | 100 |

В итоге было выбрано 256 компонент гауссовой смеси для моделирования голоса диктора.

Таким образом, как видно в таблице, модель успешно справляется с идентификацией тех дикторов, которые были в обучающей выборке.

* 1. **Определение незарегистрированных дикторов**

Помимо идентификации дикторов из обучающей выборки, система должна еще и отвергать тех дикторов, которые к этой выборке не относятся. Таких дикторов будем называть незарегистрированными.

Самым простым способом решения этой задачи является установление некоторого порога, который будет отделять зарегистрированных дикторов от незарегистрированных. Однако, такой способ не совсем точный.

Для решения этой задачи было решено использовать универсальную фоновую модель UBM (Universal Background Model) [8]. Универсальная фоновая модель – это модель гауссовых смесей, обученная не на отдельном дикторе, а на всей обучающей выборке. Таким образом, после обучения, мы получаем модель, описывающую всех зарегистрированных дикторов.

Модели отдельных дикторов обучаются на основе UBM с помощью алгоритма MAP адаптации (Maximum A-Posteriori Adaptation). Мы будем адаптировать только медианы, не затрагивая остальных параметров. Рассмотрим подробнее алгоритм адаптации:

Пусть имеется обученная по всей выборке универсальная фоновая модель , i = 1,…, M.

На основе этих параметров вычисляются параметры адаптированной модели  по следующим формулам:







Параметр r обычно берут равным 16.

В итоге процесс идентификации происходит следующим образом:

1. Извлекаем коэффициенты MFCC из входного сигнала.
2. Для полученного набора векторов ищем среди натренированных моделей ту, которая максимизирует апостериорную вероятность.
3. Пусть нашли модель . Далее рассчитываем разность . Сравниваем S(X) с некотрым пороговым значением и на основе этого сравнения принимаем решение о принятии диктора или отказе от него.
   1. **Результаты работы программы**

Как сказано ранее, был собран набор из 100 дикторов, в котором большая часть – мужчин.

Время расчета мел-частотный кепстральных коэффициентов каждого диктора в среднем составило меньше двух секунд. Для каждого диктора был получен набор из векторов длиной из 39 коэффициентов.

Модели гауссовых смесей обучались до достижения сходимости параметров для достижения наибольшей точности. Количество компонент модели было выбрано равным 256.

Для корректной, несмещенной оценки качества алгоритма, была применена перекрестная проверка (cross-validation).

Суть данного метода в том, что выборка делится на k частей, затем, последовательно алгоритм обучается на одной из k частей и тестируется на остальных k – 1 частях.

Таким образом, исходная выборка была поделена на 5 частей по 20 дикторов. При каждой итерации сначала обучается универсальная фоновая модель.

Универсальная фоновая модель для 20 дикторов в среднем обучалась за 35 минут.

Адаптация 20 моделей завершалась в среднем за 21 минуту. Таким образом, в среднем, каждая модель адаптировалась около одной минуты.

Алгоритм можно оценить с помощью популярных мер качества.

Для этого определим следующую матрицу ошибок:

|  |  |  |
| --- | --- | --- |
|  | Зарегистрированные | Незарегистрированные |
| Приняты системой | TP (true positives) = 95 | FN (false negatives) = 44 |
| Не приняты системой | FP (false positives) = 5 | TN (true negative) = 356 |

Пусть P = TP + FN, N = TN + FP, P' = TP + FP, N' = FN + TN. Тогда, на основе этих значений можно получить значения следующих мер качества:

* Точность (accuracy) 
* Полнота (recall) 
* Точность (precision) 
* F – мера 

Основное количество ошибок было совершено алгоритмом при принятии решения для незарегистрированных дикторов.

Причинами понижения точности алгоритма может быть следующее:

* Обучающая выборка была смещена в сторону мужских голосов, в результате чего универсальная фоновая модель также была смещена в эту сторону, поэтому дикторы женского пола в большей вероятности отвергались системой;
* Длина записи некоторых дикторов может быть недостаточной для корректного обучения моделей, в результате чего, незарегистрированные дикторы имели больше шансов быть принятыми;
* Каждый диктор в выборке был записан в разных условиях, с разным акустическим окружением и разным качеством оборудования, что тоже может влиять на результат.

**Выводы**

В работе были решены следующие задачи:

Были изучены основные существующие методы идентификации диктора по голосу, рассмотрена их основные преимущества и недостатки.

Взвесив все плюсы и минусы различных методов, был выбран и реализован эффективный алгоритм идентификации на основе мел-частотных кепстральных коэффициентов и модели гауссовых смесей.

Была проведена экспериментальная оценка качества алгоритма, которая составляет 93% правильных решений. Выявлено, что необходимо изучить способы понижения влияния большой разности полов на результаты экспериментов. Так же необходимо адаптировать алгоритм к разным условиям записи дикторов и различной длине речевых фрагментов.

**Заключение**

Задача идентификации диктора по голосу является актуальной в настоящее время, так как имеет широкое применение в различных областях. В данной работе изучены основные методы решения этой задачи. Реализована система идентификации диктора на основе мел-частотных кепстральных коэффициентов и моделей гауссовых смесей.

Реализованный метод имеет высокий, но недостаточный уровень точности и она совершенно не гибка к добавлению новых дикторов, так как необходимо заново переобучать всю систему.

В дальнейшем планируется рассмотрение других методов как в извлечении голосовых коэффициентов, например, метод i-векторов, так и алгоритмов идентификации, например, использование совместной GMM-SVM моделей и нейронных сетей.

**Список литературы**

1. F. Z. Chelali, K. Sadeddine, A. Djeradi Speaker identification system using LPC - Application on Berber language // HDSKD international journal. 2015. Vol. 01, №. 02. С. 29 - 46.
2. S. Davis, P. Mermelstein Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. // IEEE transactions on acoustics, speech, and signal processing. 1980. №28, 4. С. 357 - 366.
3. S. Furui Cepstral analysis technique for automatic speaker verification // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1981. №29, 2. С. 254–272.
4. D. Ramage Hidden Markov Models Fundamentals // CS229 Section Notes. 2007.
5. A. Solomonoff, C. Quillen, W. Campbell Channel compensation for svm speaker recognition // ODYSSEY04 -- The Speaker and Language Recognition Workshop Toledo. 2004.
6. Иванов И. И. Анализ метода мел-частотных кепстральных коэффициентов применительно к процедуре голосовой аутентификации // Актуальные проблемы гуманитарных и естественных наук. 2015. №10-1. С. 106-114.
7. G. Saha, S. Chakroborty, S. Senapati A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications. 2005.
8. Jeff A. Bilmes A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models // International Computer Science Institute Berkeley CA. 1998. С. 7 - 13.
9. D. A. Reynolds, T. F. Quatieri, R. B. Dunn Speaker Verification using adapted Gaussian mixture models // Digital Signal Processing. 2000. Vol. 10. С. 19 - 41.