МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ
Кафедра интеллектуальных систем

ПОЛИТАЙ
Константин Дмитриевич

РАСПОЗНАВАНИЕ ЭМОЦИЙ В РЕЧЕВОМ СИГНАЛЕ

Дипломная работа

Научный руководитель:
старший преподаватель
Щетько Николай Николаевич

Допущена к защите

«___» _____ 2018 г.

Зав. кафедрой интеллектуальных систем,
кандидат физико-математических наук,
доцент К.В. Козадаев

Минск, 2018

ОГЛАВЛЕНИЕ

ОГЛАВЛЕНИЕ 2
РЕФЕРАТ 4
ВВЕДЕНИЕ 7
ГЛАВА 1 ХАРАКТЕРИСТИКА СИСТЕМ РАСПОЗНАВАНИЯ ЭМОЦИЙ 9
1.1 Модели эмоций 9
1.2 Обработка речевых сигналов 11
1.3 Влияние эмоций на характеристики речи 13
1.4 Машинное обучение 14
1.5 Особенности задачи 17
1.6 Базы данных эмоциональной речи 17
1.7 Зависимость выражения эмоций в речи от языка 19
ГЛАВА 2 ЭТАПЫ РЕШЕНИЯ ЗАДАЧИ РАСПОЗНАВАНИЯ ЭМОЦИЙ В РЕЧИ 20
2.1 Предварительная обработка 20
2.2 Выделение признаков 20
2.2.1 Высота звука 22
2.2.2 Мел-частотные кепстральные коэффициенты 22
2.2.3 Преобразование Фурье 24
2.2.4 Кодирование с линейным предсказанием 24
2.2.5 Вейвлет-преобразования 25
2.2.6 Декомпозиция на эмпирические моды и преобразование Гильберта – Хуанга 27
2.2.7 Анализ с использованием корреляционной функции 28
2.2 Отбор признаков 28
2.3 Классификация 29
2.3.1 Метод ближайших соседей 30

2.3.2 Метод опорных векторов 31
2.3.3 Модель гауссовых смесей 32
2.3.4 Скрытые марковские модели 32
2.3.5 Нейросетевые методы классификации 34
2.3.6 Анализ с использованием динамического трансформирования времени 38
2.4 Оценка результатов обучения 39
2.4.1 Переобучение 39
2.4.2 Скользящий контроль 39
ГЛАВА 3 РЕАЛИЗАЦИЯ СИСТЕМЫ РАСПОЗНАВАНИЯ ЭМОЦИЙ .. 41
3.1 Используемые инструменты 41
3.2 Описание системы 41
3.3 Результаты распознавания 44
ЗАКЛЮЧЕНИЕ 48
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 49

РЕФЕРАТ

Дипломная работа: 54 страницы, 10 рисунков, 4 таблицы, 18 использованных источников, 1 приложение.

РАСПОЗНАВАНИЕ ЭМОЦИЙ, ОБРАБОТКА РЕЧИ, МАШИННОЕ ОБУЧЕНИЕ, КЕПСТРАЛЬНЫЙ АНАЛИЗ.

Объект исследования – записи эмоциональной речи.

Цель работы – разработка метода определения эмоциональной составляющей в речевом сигнале.

Методы исследования – компьютерное моделирование.

В исследовании описывается связь эмоционального состояния человека и параметров речевого сигнала, а также рассматриваются различные варианты реализации системы автоматического распознавания эмоций в речи.

В результате проведенного исследования были реализованы системы, использующие в качестве входных данных мел-частотные кепстральные коэффициенты и различные алгоритмы классификации.

Также в работе были проанализированы результаты распознавания эмоций для контрольных данных и проведен сравнительный анализ эффективности и точности различных подходов. Было установлено, что точность автоматического распознавания эмоций из речи может достигать значений, сравнимых с точность распознавания эмоций человеком.

Результаты работы могут быть использованы при разработке систем автоматического распознавания эмоций для выбора ее конфигурации или параметров.

РЭФЕРАТ

Дыпломная праца: 54 старонкі, 10 малюнкаў, 4 табліцы, 18 выкарыстанных крыніц, 1 дадатак.

РАСПАЗНАВАННЕ ЭМОЦЫЙ, АПРАЦОЎКА ГАВОРКІ, МАШЫННАЕ НАВУЧАННЕ, КЕПСТРАЛЬНЫЙ АНАЛІЗ.

Аб'ект даследавання – запісы эмацыйнай гаворкі.

Мэта – распрацоўка метаду вызначэння эмацыйнай састаўляючай у галасавым сігнале.

Метады даследавання – камп’ютарнае мадэляванне.

У даследаванні апісваецца сувязь эмацыйнага стану чалавека і параметраў маўленчага сігналу, а таксама разглядаюцца розныя варыянты рэалізацыі сістэмы аўтаматычнага распазнавання эмоцый у гаворцы.

У выніку праведзенага даследавання былі рэалізаваны сістэмы, якія выкарыстоўваюць у якасці ўваходных дадзеных мел-частотныя кепстральныя каэфіцыенты і розныя алгарытмы класіфікацыі.

Таксама ў працы былі прааналізаваныя вынікі распазнавання эмоцый для кантрольных дадзеных і праведзены параўнальны аналіз эфектыўнасці і дакладнасці розных падыходаў. Было ўстаноўлена, што дакладнасць аўтаматычнага распазнавання эмоцый з гаворкі можа дасягаць значэнняў, параўнальных з дакладнасцю распазнання эмоцый чалавекам.

Вынікі работы могуць быць выкарыстаны пры распрацоўцы сістэм аўтаматычнага распазнавання эмоцый для выбару яе канфігурацыі або параметраў.

ABSTRACT

Thesis: 54 pages, 10 figures, 4 tables, 18 sources, 1 application.

EMOTION RECOGNITION, SPEECH PROCESSING, MACHINE LEARNING, CEPSTRAL ANALYSIS.

The object of research – records of emotional speech.

Objective – development of a method for determining the emotional component in a speech signal.

The methods – computer simulation.

The study describes the relationship between the emotional state of a person and the parameters of a speech signal, as well as various approaches for implementing a system for automatic emotion recognition from speech.

As a result of the research, different systems were implemented that use mel-frequency cepstral coefficients as input data and various algorithms for classification.

Also in the work were analyzed the results of the emotion recognition for test data and made a comparative analysis of the effectiveness and accuracy of different approaches. It was found that the accuracy of automatic emotion recognition from speech can reach values comparable to the accuracy of human emotion recognition.

The results of the work can be used in developing automatic emotion recognition systems for choosing its configuration or parameters.

ВВЕДЕНИЕ

Несмотря на то, что распознавание эмоций из речевого сигнала является относительно новой областью исследований, у систем автоматического распознавания эмоций из речи существует большое количество возможных применений. Они могут применяться во взаимодействиях между людьми и между людьми и машинами, системы распознавания эмоций могут помогать предоставлять пользователям услуги, адаптированные к их эмоциям.

С развитием технологий увеличивается интерес в создании машин, имеющих поведение, похожее на человеческое. Также удовлетворенность пользователей может быть определена с помощью систем распознавания эмоций. Помимо этого, данные системы могут быть использованы для определения того, находится ли человека в состоянии гнева или разочарования. В случаях, когда люди находятся в негативных эмоциональных состояниях, могут быть не допущены к выполнению ответственных действий, например, вождению автомобиля, пилотированию самолета или проведению операции.

Также можно представить систему, автоматически сопровождающую разгневанных посетителей в колл-центр к оператору или мощную поисковую машину, которая ищет людей, которые обсуждают определенную проблему в определенном эмоциональном состоянии. С научной точки зрения проблема автоматического распознавания эмоций из речи интересна тем, что она преодолевает пропасть между низкоуровневыми признаками речевого сигнала и высокоуровневой и даже субъективной информацией об эмоциях.

Передача информации от человека к человеку осуществляется во многом через речь. Следовательно, акустическая часть голосового сигнала содержит значительную часть информации об эмоциональном состоянии человека.

Эмоциональная информация, находящаяся в речи, является важным фактором коммуникации и взаимодействия между людьми, так как обеспечивает обратную связь в общении, несмотря на то, что не меняет содержание речи. Информацию, передаваемую в разговорной речи, можно разделить на два типа: основной канал, связанный с синтаксической и семантической частью, который передает лингвистическую информацию, и вторичный канал, который передает паралингвистическую информацию, такую как тон, эмоциональное состояние и жесты.

Анализ эмоций в речи основывается на методах выделения эмоциональных реплик как маркеров эмоционального состояния, определенного настроения или стресса. Основное предположение состоит в

том, что произносимые реплики могут быть использованы для предсказания эмоционального состояния говорящего. Это предположение обоснованно тем, что переживание эмоций вызывает психологические реакции, которые влияют на процесс формирования речи. Например, в состоянии страха обычно повышается частота сердцебиения, учащается дыхание, появляется потливость и мышечное напряжение. В результате этого изменяется вибрация голосовых складок и форма речевого тракта. Все это влияет на акустические характеристики речи, которые позволяют слушателю распознать эмоции собеседника.

Эмоциональный искусственный интеллект – изучение и разработка систем и устройств распознавания, интерпретации, обработки и имитации человеческих аффектов. Это междисциплинарная область, охватывающая программирование, психологию и когнитивистику. Происхождение данного направления можно найти еще в ранних философских исследованиях эмоций, однако современное направление берет начало в статье по аффективным вычислениям Розалинды Пикард, написанной в 1995 году. Предпосылкой для этих исследований стало исследование возможности имитации сопереживания. Искусственные системы смогут интерпретировать эмоциональное состояние людей и адаптировать свое поведение к ним, давая соответствующую реакцию на эти эмоции [19].

ГЛАВА 1

ХАРАКТЕРИСТИКА СИСТЕМ РАСПОЗНАВАНИЯ ЭМОЦИЙ

1.1 Модели эмоций

Эмоции играют важнейшую роль во взаимодействии людей с внешним миром и имеют огромное влияние на процесс принятия решений людьми. Поэтому определение эмоционального состояния является ключевым фактором в общении. Каждая эмоция имеет уникальные свойства, которые позволяют ее распознать. Акустический сигнал, полученный при произнесении одной и той же фразы, изменяется в первую очередь из-за биологических изменений (таких как вызванное стрессом сужение гортани), вызванных эмоциями [3].

Эмоциональный интеллект – способность чувствовать, выражать и управлять собственными эмоциями, а также воспринимать и распознавать эмоции других. В психологии эмоциональное состояние определяется как сложное состояние, которое приводит к психологическим и физиологическим изменениям, которые влияют на поведение и мышление.

Для того чтобы провести классификацию эмоций, выделяется множество различных их типов. Существует большое количество моделей эмоций, созданных различными исследователями в различное время. В некоторых из этих моделей определенные эмоции выделяются как основные, или первичные. Некоторые определяют модели смешанных эмоций, вместо разделения их. Поскольку интерес к автоматическому распознаванию эмоций возрастает, выявление множества рассматриваемых эмоций становится более важным.

Одной из теорий является теория палитры, предложенная Декартом в 1956 году с целью описания всех эмоций как смеси базовых эмоций [4]. Базовыми эмоциями в ней предполагаются злость, отвращение, страх, веселье, грусть и удивление. Эти эмоции являются наиболее отчетливыми. Другой набор базовых эмоций был предложен Экманом в 1992 году [4]. Предложенными эмоциями были злость, счастье, удивление, отвращение, грусть и страх. Эта теория основывается на рассмотрении выражений лица для данных базовых эмоций как уникальных и универсальных. Колесо эмоций – еще одна популярная теория [11]. Колесо эмоций состоит из 8 базовых эмоций и 8 составных эмоций, каждая из которых является сочетанием двух основных. К основным эмоциям относится веселье, доверие, страх, удивление, грусть, отвращение, злость и ожидание. На рисунке 1.1 изображена модель колеса эмоций Плутчика.

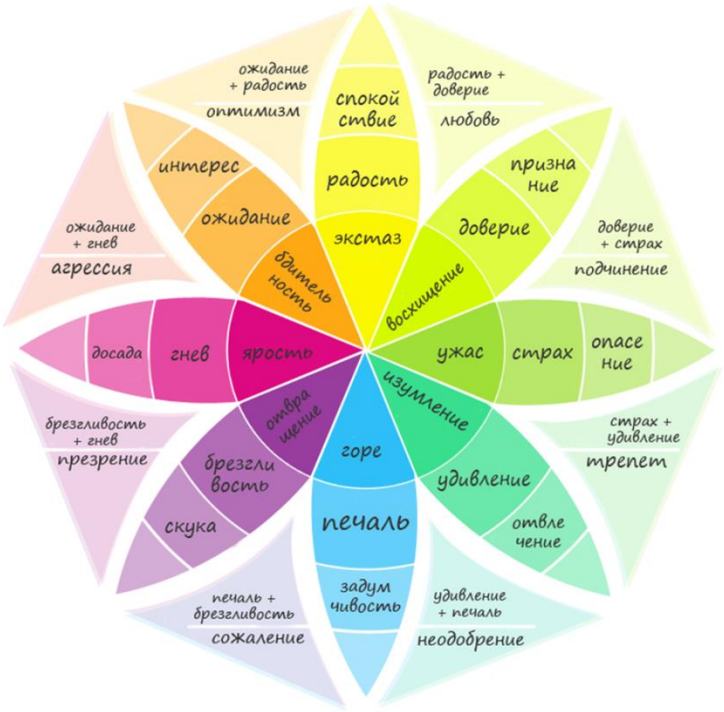

Рисунок 1.1 – Модель колеса эмоций, предложенная Плутчиком

Эмоции в звуковом сигнале могут быть описаны как с помощью отдельных описаний каждой из эмоций, так и представлены в виде эмоционального пространства определенной размерности [10].

В последние годы становится популярной постановка задачи, называемая теорией эмоциональных размерностей. Считается, что любую речь можно описать с помощью двух или трех размерных параметров. В англоязычной литературе часто используется VAD-модель [7], в которой используются метрики valence (позитивная-негативная), arousal (неожиданность) и dominance (контролируемость), которые меняются на протяжении всего высказывания. Полагается, что в трехмерном VAD-пространстве образуются области, которые соответствуют конкретным эмоциям. Таким образом, в данном подходе, все эмоции разделяются на категории по трем основным параметрам. Основное преимущество предсказания при помощи размерных метрик эмоциональной речи состоит в большей информативности по сравнению с описанием дискретными классами. Недостатком данного подхода является субъективность, так как их оценка зависит от различных факторов, в том числе случайных. По этой причине намного сложнее получить размеченную базу.

На самом деле в акустических признаках содержится не вся информация о высказывании. Это иллюстрирует психологический эксперимент Мак-Гурка проведенный в 1976 году [8]. В нем показано, что человек при восприятии речи использует не только звуковые данные, но и визуальные. Это позволяет утверждать, что только акустических признаков может быть недостаточно для полноценного автоматического анализа речи, в том числе распознавания эмоций.

Люди общаются между собой при помощи речи на протяжении многих тысячелетий. В последнее время взаимодействие между людьми и машинами стало быстро развивающейся областью в промышленности и в академических исследованиях. Речь является одним из фундаментальных способов взаимодействия между людьми. Речевой сигнал является логически упорядоченной последовательностью звуков. Человеческий мозг производит сложный набор операций для анализа поступающей звуковой информации, в том числе речи. Он преобразует звуки в некоторые концептуальные идеи и мысли, которые образуют набор инструкций, команд или информацию.

Автоматическое распознавание часто используется для определения эмоций из определенного заданного набора их классов. Распознавание эмоций из речи является видом распознавания голоса. Обработка речи в этом случае проходит три главных этапа, таких как предварительная обработка, выделение признаков и определение закономерностей. В случае речевого сигнала, гласные содержат в себе наибольшую информативную часть. Гласные являются наиболее сильно произносимыми голосом частями слов. Отсюда следует, что является желательным выделение части сигнала, произносимой голосом, и дальнейшая обработка только этой части.

Для эффективного и естественного взаимодействия между людьми и машинами распознавание эмоций играет очень важную роль. Эмоции отображают психическое состояние людей через речь, выражение лица, позы и жесты, а также другие физиологические параметры, такие как температура тела, кровяное давление, напряжение мышц и другие. Психологическое состояние человека косвенно влияет на речь, произносимую человеком. Например, при взаимодействии между людьми скорость речи увеличивается, если человек злится или радуется, а диапазон высоты тона становится шире, если же человек находится в состоянии грусти, то речь замедляется, а диапазон изменения высоты тона становится меньше. Распознавание эмоций в речи имеет ряд преимуществ в различных приложениях [17].

1.2 Обработка речевых сигналов

Обработка речи – область научных исследований, в которой осуществляются фильтрация, усиление и извлечение информации, кодирование, сжатие и восстановление речевых сигналов. Обработка речи в системах распознавания применяется для решения следующих задач:

1) фильтрация речи и шумоподавление;
2) разделение речи на информативные сегменты;

3) определение информативных параметров;
4) распознавание речи.

Каждая задача обработки речевого сигнала реализуется с использованием определенных методов. В зависимости от области, в которой производится обработка, все методы можно разделить на три области: частотную, временную и частотно-временную.

Методы обработки во временной области заключаются в определении характерных точек речевого сигнала с дальнейшим использованием их для анализа. С точки зрения технической реализации в качестве характерных точек выбираются явные максимумы (минимумы) или моменты пересечения нулевой оси времени функцией речевого сигнала. Главным недостатком методов обработки речевого сигнала во временной области является неоднозначность выделения характерных точек, которая вызывается шумами и смещениями нулевого уровня.

Методы обработки в частотной области состоят в использовании всех отсчетов данных, имеющихся в речевом сигнале. Многие речевые сигналы имеют характерный специфический частотный состав и занимают определенные спектральные области. Использование таких методов позволяет производить обработку речевых сигналов с достаточно высокой точностью. К недостаткам обработки в частотной области можно отнести низкую адаптивность к локальным свойствам сигналов, недостаточно высокое спектральное разрешение и сравнительно большие затраты.

Методы обработки в частотно-временной области представляют собой методы, которые включают все преимущества временного и частотного анализов с меньшими проявлениями их недостатков.

Анализ существующих методов обработки речевых сигналов показывает, что в зависимости от характера обработки методы можно разделить на группы, реализующие различные виды анализа [1]:

1) с использованием преобразования Фурье;
2) с использованием вейвлет-преобразования;
3) с использованием декомпозиции на эмпирические моды и преобразования Гильберта – Хуанга;
4) с использованием кепстра (кепстральный анализ);
5) с использованием линейного предсказания;
6) с использованием корреляционной функции (корреляционный анализ;
7) с использованием нейронных сетей;
8) с использованием скрытых марковских моделей;
9) с использованием динамического трансформирования времени.

1.3 Влияние эмоций на характеристики речи

Систематические исследования выражения эмоций с помощью голоса были начаты еще в XIX веке различными естествоиспытателями, учеными, наблюдающими за поведением человека и животных.

Исследователями установлен тот факт, что человеческая речь, произносимая им в различных эмоциональных состояниях, различается по целому ряду признаков. К числу наиболее информативных признаков относятся, в первую очередь, характеристики просодической группы, которые точно отражают процессуальную сторону устных высказываний и, прежде всего, изменяются при реакциях аффективного плана.

Задача автоматического распознавания речевого сигнала и, в частности, ее эмоциональной окраски является междисциплинарной и привлекает исследователей различных специальностей – не только лингвистов, но и математиков, программистов, психологов, физиологов. От решения данной задачи зависит прогресс современных автоматизированных систем управления, реабилитации и протезирования, систем безопасности, срочного оповещения. Решение задачи распознавания речи имеет большое научное значение для многих сфер фундаментальных исследований человека и информационных технологий. За последнее время существенно усилился интерес к анализу речевых сигналов, рассматриваемых в качестве наиболее удобного объективного показателя выражаемых эмоций и эмоционального состояния человека.

Данные рассуждения касаются не только сфер деятельности с повышенной ответственность, таких как космонавтика, авиация (летчики, диспетчеры аэропорта), обслуживание АЭС и других, которые изначально доминировали в этом отношении, но также и бытовой сферы.

Несмотря на то, что изучение эмоций точными научными методами еще только начинается, уже сейчас стало очевидным большое значение этой проблемы, как для теоретической науки, так и для практического применения. При этом понятно, что решить эту задачу нельзя без знания алфавита акустического языка эмоций. Однако для этого необходимо выделить признаки, ответственные за формирование эмоциональности голоса.

Несмотря на большое количество разнообразных исследований и коммерческих решений в области распознавания эмоций в речи, проблема автоматического распознавания эмоционального состояния человека на сегодняшний день не является до конца решенной, в частности, отсутствует модель описания речевых образцов в условиях проявления различных видов эмоций. Процесс интерпретации (распознавания) эмоций человека по

естественной речи является сложной задачей, как в области математической формализации, так и в смысле поиска способов четкой конкретизации эмоционального состояния – однозначного детектирования эмоции по характеристикам речевого сигнала.

Отсутствие универсальной теоретической модели описания эмоций обусловлено целым рядом взаимосвязанных проблем. С одной стороны, необходимо выделить в речи те параметры, которые могли бы служить для индикации эмоций. Здесь возникают проблемы их регистрации, математического анализа, поиска соответствующих алгоритмов и технических средств. Для решения данной задачи необходимо четко задать «входные» и «выходные» данные, формально представить желаемый результат. С другой стороны, необходимы формальные, объективные методы, чтобы систематизировать и классифицировать такие сложные явления, как эмоции человека. Необходимо разработать адекватную модель и собрать базу данных – набор определенных «образцов» состояний и соответствующих им записей речи. Получается замкнутый круг: чтобы решить одну задачу, надо уже иметь решение другой.

Однако научные исследования и практические разработки в этом направлении предпринимаются с возрастающей интенсивностью, подстегиваемой коммерческими возможностями. При этом, чаще всего, разработчики новых методов и инструментов анализа пользуются только собственным «здравым смыслом» и некоторыми теоретическими обобщениями психологов и фонологов. Последним же для анализа эмоциональных явлений приходится пользоваться «стандартными», общедоступными инструментами объективного анализа речевых сигналов. Чтобы как-то приблизиться к достижению практической эффективности, всем необходимо упрощать задачу – при разработке новых методов анализа речевого сигнала ограничиваться только отдельными аспектами эмоциональных феноменов, например, лишь интерпретацией знака эмоций или отдельных эмоций, наиболее важных для конкретной области применения.

1.4 Машинное обучение

С развитием технологий и ростом таких областей, как машинное обучение, обработка звука и речи, эмоциональные состояния становятся неотъемлемой частью взаимодействия людей и машин. Все больше исследований посвящается исследованию возможности создания машин, позволяющих распознавать, интерпретировать и симулировать эмоции.

Машинное обучение — подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Одним из типов обучения является обучение по прецедентам, которое основано на выявлении общих закономерностей по частным эмпирическим данным. В общей постановке данной задачи дано конечное множество прецедентов (объектов, ситуаций), по каждому из которых собраны (измерены) некоторые данные. Данные о прецеденте называют также его описанием. Совокупность всех имеющихся описаний прецедентов называется обучающей выборкой. Требуется по этим частным данным выявить общие зависимости, закономерности, взаимосвязи, присущие не только этой конкретной выборке, но вообще всем прецедентам, в том числе тем, которые ещё не наблюдались.

Наиболее распространённым способом описания прецедентов является признаковое описание. Фиксируется совокупность n показателей, измеряемых у всех прецедентов. Если все n показателей числовые, то признаковые описания представляют собой числовые векторы размерности n. Возможны и более сложные случаи, когда прецеденты описываются временными рядами или сигналами, изображениями, видеорядами, текстами, попарными отношениями сходства или интенсивности взаимодействия, и т. д.

Задача автоматического распознавания может быть решена как задача машинного обучения.

Обучение с учителем — наиболее распространённый случай типов задач машинного обучения. Каждый прецедент представляет собой пару «объект, ответ». Требуется найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ. Функционал качества обычно определяется как средняя ошибка ответов, выданных алгоритмом, по всем объектам выборки.

Классификация — один из разделов машинного обучения, посвященный решению следующей задачи. Имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

Классифицировать объект — значит, указать номер (или наименование класса), к которому относится данный объект. Задача классификации отличается тем, что множество допустимых ответов конечно. Их называют метками классов. Класс — это множество всех объектов с данным значением метки. В машинном обучении задача классификации относится к разделу обучения с учителем.

Если имеется фиксированное число классов эмоций, которые необходимо распознать, и обучающая выборка, для которой определена принадлежность каждой записи к определенному классу эмоций, то задача автоматического распознавания эмоций является задачей классификации.

На рисунке 1.2 представлен процесс решения задачи автоматического распознавания эмоций. Система распознавания эмоций из речи состоит из пяти основных компонент: входной базы данных записей эмоциональной речи, выделения признаков, отбора признаков, классификации и вывода распознанного результата. В целом система основывается на глубоком анализе механизма генерации речевого сигнала, выделении отдельных признаков, содержащих информацию об эмоциях говорящего и учета необходимых закономерностей модели распознавания для идентификации типа эмоций.

Вначале используется база данных эмоциональной речи, которая содержит записанные и помеченные фразы. Затем производится выделение признаков из исходных данных. После этого происходит отбор наиболее значимых и релевантных признаков для снижения размерности модели. Далее на основе полученных данных производится обучение классифицирующего алгоритма.

Рисунок 1.2 – Процесс решения задачи автоматического распознавания эмоций в машинном обучении.

Обычный набор человеческих эмоций имеет около 300 эмоциональных состояний. Всякий раз сигнал проходит через выделение признаков и процесс

их отбора, причем признаки выбираются исходя из их соответствия задаче распознавания эмоций. Первой ступенью является генерация базы данных для обучения и проверки выделенных признаков. В конце процесс распознавания эмоций из речи завершается применением классификатора [17].

Распознавание эмоций из речи напоминает задачу распознавания говорящего, однако они имеют различные подходы для определения эмоций и для того, чтобы сделать этот процесс надежным и точным. Качество распознавания полученной системы зависит от естественности выходных данных.

Многочисленные алгоритмы машинного обучения могут быть применены для осуществления классификации. После процедуры выделения признаков, необходимо произвести классификацию. Целью построения модели классификации является предсказание при помощи алгоритмов машинного обучения эмоционального состояния на основе параметров речи.

1.5 Особенности задачи

До сих пор существует большое количество не до конца ясных моментов при выборе алгоритма классификации эмоций. Различные комбинации эмоциональных признаков дают разную степень точности классификации эмоций. Исследователи до сих пор спорят, какие из признаков оказывают наибольшее влияние на распознавание эмоций [17].

Распознавание эмоций из речи является сложной задачей по ряду причин:

1) неясно, какие признаки наиболее эффективны для различения эмоций в речи;
2) акустическая изменчивость, обусловленная различием в произносимых фразах, дикторах и различной скоростью речи напрямую влияет на большинство извлекаемых речевых признаков, таких как высота звука и энергия;
3) распознавание эмоций людьми происходит по выражению лица говорящего, его голосу и содержанию его высказывания в совокупности.

1.6 Базы данных эмоциональной речи

Корпус записей речи для изучения эмоций должен представлять различные типы эмоций с различными конфигурациями.

Существует большое количество баз данных эмоциональной речи на различных языках: немецком, английском, японском, испанском, китайском, русском и т. д. Одна из основных характеристик базы данных эмоциональной речи – тип выражаемых эмоций. Они могут быть симулированными или взятыми из реальных жизненных ситуаций. Преимуществом симулированных записей эмоций является то, что исследователь имеет полный контроль над качеством записи аудио. Однако их недостатком является меньший уровень естественности и спонтанности. Не симулированные базы данных состоят из записей речи, взятых из реальных жизненных ситуаций, таких как звонки в колл-центры, интервью, заседания, фильмы, видеозаписи и другие подобные, где проявляются спонтанные и естественные эмоции. Недостатком является неполный контроль над проявляемыми эмоциями. Также проблемой является возможное низкое качество записи.

Одной из известных и часто используемых баз является Берлинская база данных эмоциональной речи. База данных содержит записи шести основных эмоций (злость, скука, отвращение, тревога, счастье и грусть), а также нейтральной речи. Профессиональные немецкие дикторы (5 мужчин и 5 женщин) симулировали данные эмоции, записав 10 фраз (5 коротких и 5 длинных предложений), которые используются в повседневном общении и могут быть интерпретированы со всеми данными эмоциями. Записанный речевой материал, содержащий около 800 записей, был оценен по узнаваемости и естественности при прослушивании 20 волонтерами. Волонтерам требовалось распознать и оценить естественность выражаемых эмоций при прослушивании случайных записей. Записи, которые более 60% судей оценили как естественные и 80% правильно распознали, были включены в итоговую базу данных. После отбора в базе данных осталось 535 предложений. Предложения не равномерно распределены по различным эмоциональным состояниям.

Таблица 1.1 – Состав Берлинской базы данных эмоциональной речи.

Эмоция	Количество записей	Процент узнаваемости людьми, %
Злость	127	96,2
Нейтральная	79	88,2
Страх	69	87,3
Скука	81	86,2
Счастье	71	83,7
Грусть	62	80,7
Отвращение	46	79,6

1.7 Зависимость выражения эмоций в речи от языка

Распознавание эмоций может быть независимым от языка и национальности. Корреляция между акустическими признаками и основными эмоциями у различных народов достаточно похожа из-за универсальных физиологических и психологических реакций на эмоции [18].

В ряде работ была доказана относительная независимость точности распознавания эмоций от языка. Например, в работе [8] исследователи просили японских слушателей описать эмоции, выражаемые японцами и американцами в бессмысленных фразах без семантической информации. Точность распознавания эмоций людьми составила около 60%. Слушатели должны были отнести эмоции к одному из девяти заданных типов. Подобный результат был получен при автоматическом распознавании эмоций.

Данные исследования показывают, что некоторые достоверные акустические параметры предоставляют возможность добиться автоматического распознавания эмоций по голосу. С другой стороны, так как точность распознавания эмоций людьми не слишком высока, то, вероятно, не следует ожидать абсолютной точности в автоматическом распознавании эмоций. Относительно низкая точность распознавания эмоций людьми может объясняться похожими психологическими свойствами определенных эмоциональных состояний, что приводит к похожим акустическим признакам. В то время, как люди могут использовать другую контекстную информацию (жесты, поза, выражение лица и другие) для разрешения неопределенности, системы автоматического распознавания эмоций, основанные на обработке речевого сигнала, должны концентрироваться на распознавании определенных эмоциональных состояний для большей эффективности.

ГЛАВА 2

ЭТАПЫ РЕШЕНИЯ ЗАДАЧИ РАСПОЗНАВАНИЯ ЭМОЦИЙ В РЕЧИ

Рассмотрим действия, которые выполняются на каждом из этапов решения задачи классификации при разработке системы автоматического распознавания эмоциональной речи.

2.1 Предварительная обработка

Предварительная обработка – это операции, которые необходимо выполнить к записям речевых сигналов перед выделением признаков. Например, из-за различия в условиях записи, требуется выполнить нормализацию энергии каждой записи высказывания. Из всей записи удаляются части, содержащие молчание и не несущие никакой информации. Далее производится оценка энергии сигнала и ее нормализация. Затем сигнал очищается от шума путем задания порогового значения коэффициентов.

Для более успешного распознавания эмоционального состояния человека по голосу предварительно может определяться пол говорящего, так как было показано, что одни и те же эмоции у мужчин и женщин могут иметь различные значения признаков.

2.2 Выделение признаков

Одним из первых шагов в любой системе автоматического распознавания речи является выделение признаков, то есть идентификация компонентов звукового сигнала, которые позволяют хорошо распознать содержимое, и отбрасывание других, содержащих шум и постороннюю информацию.

Для конструктивного решения задачи автоматического распознавания эмоций по речи необходимо количественно охарактеризовать речевой сигнал и выделить существенные параметры, отвечающие за эмоции человека.

Выделение новых, по возможности родственных человеческому восприятию информативных признаков является одной из важнейших задач распознавания эмоций по речи. Различные исследования в области акустики, психолингвистики и психофизиологии позволили собрать сведения о множестве акустических, просодических и лингвистических характеристик речи, которые можно использовать в качестве информативных признаков при распознавании эмоционального состояния, и проявляющихся на уровне сегментов, фонем, слогов, слов и фраз [5].

Выделение признаков играет важную роль в точности системы распознавания речи и является ключевой стадией, так как выделение подходящих признаков позволяет улучшить качество классификации.

Чаще всего используются следующие признаки речевого сигнала: спектрально-временные, амплитудно-частотные, вейвлет, кепстральные и характеристики нелинейной динамики.

Спектрально-временные признаки могут быть разделены на 4 категории: непрерывные, качественные, спектральные и признаки, основанные на операторе энергии Тигера [2].

Традиционные речевые признаки, относящиеся к частоте основного тона, формантам и энергии часто использовались в предыдущих работах. Так как эти признаки изначально применялись для распознавания речи или идентификации говорящего, некоторые специфические свойства, выраженные в речи, такие как тембр или ритм речи, не учитываются в них, что требует поиска новых признаков, которые могут быть использованы на практике [18].

Наиболее часто применяемыми признаками являются признаки, связанные с фундаментальной частотой, формантами, интенсивностью, мел-частотными кепстральными коэффициентами. Согласно с исследованиями, такие статистические характеристики, как математическое ожидание, максимальное и минимальное значение и дисперсия оригинальных акустических характеристик речи имеют значение для решения задачи распознавания эмоций.

Признаки, выделяемые из речи, можно сгруппировать на частотные характеристики, энергетические характеристики и мел-частотные кепстральные коэффициенты. Частотные характеристики включают в себя статистические характеристики фундаментальной частоты и первых трех формант.

Диапазон высоты тона расположен между 60 Гц и 450 Гц для звонких. Фундаментальная частота и форманты рассчитываются для окон размером 20 мс с перекрытием 10 мс, потому что для них речевой сигнал может считаться стационарным во временной шкале и статистические характеристики фундаментальной частоты и формант на длине речевого сегмента могут быть использованы как признаки. Фундаментальная частота рассчитывается по

методу автокорреляции, а форманты могут быть найдены как корни полинома кодирования с линейным предсказанием. Фундаментальная частота и форманты рассчитываются на участках, на которых произносятся гласные. Для согласных фундаментальная частота и форманты подразумеваются нулевыми и не учитываются в статистике [18].

К энергетическим характеристикам речевого сигнала относится распределение энергии по спектру, особенно для низкочастотной энергии (до 250 Гц).

Энергетические параметры также рассчитываются для окон шириной 20 мс с перекрытием 10 мс. Длительность энергетических плато приблизительно показывает длительность гласных, а длительность энергетических впадин показывает приблизительную длительность периодов тишины во фразах. Неравномерности в энергетических характеристиках показывают интонацию. Следовательно, данные характеристики могут быть использованы для выделения и анализа содержащихся в речи эмоций [18].

2.2.1 Высота звука

Высота – основная частота гортанного возбуждения. Высота зависит от напряжения вокальных складок и давления воздуха в них. Частота основного тона является одним из самых широко используемых в системах автоматического распознавания речи признаков. Время, прошедшее между последовательными открытиями определяет частоту основного тона. Из частоты основного тона могут быть извлечены такие признаки, как минимальное и максимальное значение, математическое ожидание, среднеквадратичное отклонение, коэффициент регрессии и его среднеквадратичное отклонение и др. [16]

2.2.2 Мел-частотные кепстральные коэффициенты

В области обработки речевых сигналов кепстральный анализ имеет широкую распространенность, которую можно объяснить достоинством сжатия информации о речевом сигнале при переходе в частотную область обработки.

При преобразовании сигнала из временной области в частотную область информация оказывается более подробной и компактной. Исходя из указанных достоинств спектрального представления информации, появилась идея кепстрального анализа, состоящая в замене в спектре оси частоты на ось времени.

Таким образом, появляется возможность представить исходную спектральную информацию еще более компактно, когда каждый

гармонический ряд исходного спектра будет представлен всего одной составляющей в кепстре.

Кепстр – это спектр логарифма спектра исходного сигнала, т.е. первоначальный спектр должен быть представлен в логарифмическом масштабе.

Выделение признаков может быть реализовано с использование мел-частотных кепстральных коэффициентов. Они используются для представления речевых сигналов для различных приложений, таких как распознавание речи и идентификация по голосу [13].

Вычисление мел-частотных кепстральных коэффициентов основано на дискретном косинусном преобразовании логарифма спектральной плотности мощности в нелинейной мел-шкале частот.

Вычисление мел-частотных кепстральных коэффициентов происходит в несколько этапов, которые можно кратко описать следующим образом:

1) Сигнал разделяется на короткие фрагменты.
2) Для каждого кадра рассчитывается периодограмма спектральной плотности мощности.
3) Применяется гребенка фильтров к спектральной плотности мощности, суммируется энергия каждого фильтра.
4) Берется логарифм энергии.
5) Рассчитывается дискретное косинусное преобразование.
6) Берутся 2- 13 коэффициенты, остальные отбрасываются.

Для понимания необходимости каждого из этих этапов следует разобраться с тем, как формируется человеческая речь.

Звуковой сигнал постоянно изменяется, однако для упрощения предполагается, что на коротких временных промежутках изменения незначительны. Поэтому исходный сигнал разделяется на фрагменты (кадры) длиной 20-40 мс. Если выбрать более короткие кадры, то в них будет содержаться недостаточное количество отсчетов сигнала для получения надежной спектральной характеристики. Если размер кадра больше, то сигнал будет изменяться значительно в пределах кадра.

Следующим шагом является расчет спектральной плотности мощности для каждого кадра. Это объясняется тем, что улитка в человеческом ухе вибрирует в разных местах в зависимости от частоты звука. Периодограмма сигнала выполняет аналогичную функцию, показывая, какие частоты присутствуют в сигнале.

Периодограмма по-прежнему содержит много информации, ненужной для автоматического распознавания речи. В частности, улитка уха не определяет различие между двумя близкими частотами. Этот эффект

становится более выраженным по мере роста частоты. Это для устранения этого эффекта применяется гребенка фильтров.

Энергия каждого фильтра логарифмируется. Это также объясняется устройством человеческого слуха: воспринимаемая громкость нелинейно зависит от энергии звукового сигнала.

Последним шагом является вычисление дискретного косинусного преобразования. Это объясняется тем, что из-за частичного перекрытия фильтров между ними имеется корреляция. ДКП устраняет корреляцию, что означает возможность применения диагональной матрицы ковариации для моделирования признаков, например, в смешанной гауссовой модели.

Значения мел-частотных кепстральных коэффициентов не очень надежны при наличии аддитивного шума, поэтому для уменьшения влияния шума их значения в системах распознавания речи обычно нормализуют.

Также мел-частотные кепстральные коэффициенты учитывают индивидуальные особенности говорящего. Для создания независимой от диктора системы распознавания эмоциональной речи следует избавиться от данной особенности. Для этого применяется метод нормализации кепстрального среднего. При этом вычисляется кепстральное среднее, которое приближенно описывает характеристики канала передачи и вычитается из кепстральных коэффициентов.

2.2.3 Преобразование Фурье

Преобразование Фурье широко используется во многих областях науки и технологий, в том числе и в обработке речевых сигналов. В данной области преобразование Фурье рассматривают как преобразование сигнала из временной области в частотную область и разложение его на частотные составляющие [1].

При решении задач цифровой обработки сигналов часто используют дискретное преобразование Фурье, так как речевой сигнал часто представляется в дискретном виде как сумма гармонических составляющих.

Получение спектра с использованием дискретного преобразования Фурье позволяет компактно и наглядно представить информацию о сигнале. Однако в таком виде практически невозможно детально проанализировать кратковременные локальные особенности сигнала, что является серьезным недостатком дискретного преобразования Фурье.

2.2.4 Кодирование с линейным предсказанием

Линейное предсказание является одним из часто используемых методов в задачах обработки речевых сигналов. Модель линейного предсказания основывается на предположении, что любой отсчет речевого сигнала можно приближенно оценить линейной комбинацией некоторого числа предшествующих ему отсчетов.

Основной задачей линейного предсказания является определение набора коэффициентов предсказания, которые минимизируют ошибку представления коэффициентов [1].

Используется два основных метода определения линейного предсказания, которые называются автокорреляционным и ковариационным методами решения. Оба метода используют представление сигнала во временной области. Коэффициенты предсказания определяют частотную характеристику фильтра, характеризующего состояние голосового тракта в определенный момент времени. Данные методы вычисления обеспечивают получение некоторой средней оценки анализируемого участка сигнала в частотно-временной области.

Линейное предсказание – метод, позволяющий получить спектральную плотность мощности. Это один из самых успешных широко используемых методов и мощный метод для анализа закодированных голосовых файлов с низкой частотой дискретизации.

Данный метод имеет ряд преимуществ: он позволяет получить лучшую аппроксимацию спектральных коэффициентов; требует малого времени для вычисления параметров сигнала; позволяет получить важные характеристики сигнала.

При кодировании с линейным предсказанием значение отсчета сигнала может быть представлено как линейная комбинация предыдущих отсчетов. Записывая это для всех отсчетов сигнала, можно получить систему линейных уравнений, которая может быть решена и определяет значение коэффициентов. Эти значения позволяют получить сигнал, очищенный от шума и хорошо воспроизводящий форманты.

2.2.5 Вейвлет-преобразования

Несмотря на большую практическую распространенность преобразования Фурье, многие задачи обработки речевых сигналов реализуются с использованием вейвлет-преобразования.

По сравнению с преобразованием Фурье, вейвлет-преобразование обладает рядом существенных преимуществ, которые следуют из возможности анализа кратковременных локальных особенностей сигналов, таких как короткие всплески или провалы, разрывы и ступеньки и другие.

Вейвлет-преобразование – преобразование, переводящее сигнал из временной области в частотно-временную область, что позволяет получить дополнительную информацию об анализируемом сигнале (изменение значений частотных компонент сигнала во времени). Для решения задачи распознавания эмоционального состояния человека используется непрерывное вейвлет-преобразование, так как оно позволяет анализировать сигнал на произвольно выбираемых масштабах и частотах.

Вейвлетом (материнским вейвлетом) называется некоторая функция, хорошо локализованная (сосредоточенная в небольшой окрестности некоторой точки и резко убывающая до нуля по мере удаления от нее) как во временной, так и в частотной области. К материнскому вейвлету можно применить две операции: сдвиг (смещение области локализации во времени) и масштабирование (растяжение или сжатие области локализации по частоте) [1].

Основная идея вейвлет-преобразования заключается в разбиении исходного сигнала на масштабированные и сдвинутые по оси времени версии материнского вейвлета и вычислении коэффициентов корреляции участков сигнала и версий вейвлета на заданном масштабе. Результатом этого является набор коэффициентов, показывающих, насколько поведение сигнала в данный момент времени похоже на поведение вейвлета на данном масштабе, таким образом, коэффициенты отражают близость сигнала к вейвлету данного масштаба. Если вид анализируемого сигнала в окрестности данного момента времени ближе к виду вейвлета, то соответствующий коэффициент имеет большее абсолютное значение.

Масштаб в непрерывном вейвлет-преобразовании – величина, обратно пропорциональная анализируемой частоте для сигнала. Конкретное значение масштаба однозначно определяет значение анализируемой частоты сигнала, которое принимает то или иное значение в зависимости от вида, используемого вейвлета.

Использование сдвига и масштабирования в частотно-временной области позволяет производить анализ речевых сигналов на различных масштабах и достаточно точно определять положение их особенностей во времени. В задачах обработки сигналов наиболее часто встречаются такие материнские вейвлеты, как вейвлет Хаара, вейвлет Добеши, вейвлет «Мексиканская шляпа», вейвлет Марлета (комплексный базис) [1].

Вычисление непрерывного вейвлет-преобразования на практике невозможно, поскольку любой цифровой сигнал имеет конечное число отсчетов амплитуд. Кроме этого, непосредственное вычисление требует больших вычислительных затрат.

Информация о нестационарных звуковых сигналах может быть извлечена с помощью дискретного вейвлет-преобразования. Этот метод является

сравнительно новым и вычислительно эффективным способом выделения признаков.

Каждый коэффициент вейвлет-преобразования представляет собой комплексное число, а результатом вейвлет-преобразования является матрица вейвлет-коэффициентов. Квадрат модуля вейвлет-коэффициента пропорционален энергии сигнала данной частоты в данный момент времени. По этим значениям можно определять эмоциональную окраску фрагмента сигнала.

Возможность применения вейвлетов в задачах распознавания и обработки речи вытекает из свойств речевого сигнала. Вейвлеты, как средство многомасштабного анализа, позволяют выделять одновременно как основные характеристики сигнала, так и короткоживущие высокочастотные явления в речевом сигнале.

В отличие от традиционного преобразования Фурье, вейвлет-преобразование определено неоднозначно: каждому вейвлету соответствует свое преобразование. Это позволяет тщательнее подобрать вейвлет-функцию с хорошими свойствами частотно-временной локализации.

Существуют также психофизические соображения в пользу использования вейвлет-анализа речевого сигнала. Человеческое ухо устроено так, что при обработке звукового сигнала оно передает мозгу вейвлет-образ сигнала.

2.2.6 Декомпозиция на эмпирические моды и преобразование Гильберта – Хуанга

Для адаптивного анализа и обработки речевых сигналов при помощи вейвлет-преобразования необходимо использовать априорную информацию – выбрать функцию материнского вейвлета. Вопрос о выборе соответствующей функции вейвлета на основе характеристик анализируемого сигнала не всегда однозначен. Для решения этой проблемы используется другой метод обработки, основанный на преобразовании Гильберта – Хуанга. Главным преимуществом этого метода является высокая адаптивность, проявляющаяся в том, что базисные функции, используемые при разложении сигнала, извлекаются непосредственно из самого исходного сигнала и позволяют учесть только свойственные ему особенности [1].

Преобразование Гильберта – Хуанга состоит из двух основных этапов:

1) разложения сигнала на компоненты – декомпозиция на эмпирические моды;
2) формирования по полученным эмпирическим модам спектра Гильберта.

В результате преобразования Гильберта – Хуанга речевой сигнал представляется в частотно-энергетически-временной области, что позволяет выявить скрытые модуляции и области концентрации энергии, которые позволяют анализировать как глобальные, так и локальные свойства сигналов и требуют меньших вычислительных затрат.

2.2.7 Анализ с использованием корреляционной функции

Корреляционный анализ – это определение взаимосвязи статических двух или нескольких величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). Математической мерой корреляции двух величин служит коэффициент корреляции. Корреляционный анализ статистических данных достаточно популярен в обработке речевых сигналов. Это обусловлено двумя моментами: коэффициенты корреляции относительно просты в подсчете и их применение не требует специальной математической подготовки. Применительно к задачам обработки речевых сигналов ключевыми понятиями корреляционного анализа становятся автокорреляционная и взаимно-корреляционная функции [1].

Автокорреляционная функция определяет статистическую взаимосвязь между величинами из одного речевого сигнала, разложенного в ряд, но взятых со сдвигом.

2.2 Отбор признаков

Несмотря на интуитивное предположение, что большее количество выделяемых из сигнала признаков улучшает возможности системы классификации по распознаванию эмоций, в действительности различные исследования показывают, что это не всегда так. При уменьшении размера вектора признаков классифицируемого объекта, в системе, получающей на вход более компактные и легко интерпретируемые данные, производительность классифицирующего алгоритма и скорость работы увеличивается.

Отбор признаков – это процесс выделения подмножества подходящих признаков, используемых для построения модели. Основным предположением при использовании выделения признаков является то, что данные содержат большое количество избыточных и нерелевантных признаков. Избыточными считаются признаки, не предоставляющие больше информации, чем текущие выбранные признаки, а нерелевантными – те, которые не приносят полезной информации в любом контексте.

Если в обучающей выборке содержится большое количество данных, фильтрующие методы являются предпочтительными с точки зрения вычислительной эффективности и независимости от алгоритма обучения.

2.3 Классификация

Если имеется фиксированное число классов эмоций, которые необходимо распознать, и обучающая выборка, для которой определена принадлежность каждой записи к определенному классу эмоций, то задача автоматического распознавания эмоций является задачей классификации.

Среди всех возможных подходов наиболее часто применяются методы скрытых марковских моделей, метод опорных векторов, метод максимального правдоподобия, искусственные нейронные сети, метод ближайших соседей. Другими классификаторами, заслуживающими внимания, являются деревья решений, нечеткие классификаторы и многие другие [17].

Классификаторы принимают решение на основе закономерностей в тестовых речевых фрагментах и обучающей выборке. Метод гауссовых смесей наиболее подходит для применения к глобальным признакам, которые выделяются из обучающей выборки с фразами. Классификаторы на основе искусственных нейронных сетей имеют преимущество благодаря нелинейности границ, разделяющих эмоциональные состояния. Среди многих вариантов искусственных нейронных сетей часто используемыми и наиболее простыми являются нейронные сети прямого распространения и многослойные персептроны. Искусственные нейронные сети используют концепции акустической фонетики для распознавания закономерностей. В распознавании речи скрытые марковские модели используются для классификации последовательностей данных, представленных набором различных состояний. Также они определяют вероятности перехода между этими состояниями. Скрытые марковские модели успешно применяются для моделирования быстро изменяющейся информации и спектре речевого сигнала. Метод опорных векторов относится к линейным классификаторам и может быть интерпретирован как расширение персептрона. Метод опорных векторов минимизирует эмпирическую ошибку классификации. Основной его идеей является использование различных функций в качестве ядра [17].

Различные исследователи используют различный набор эмоций в своих работах. Обычно выделяют шесть типов эмоций. Кроме того, не всегда ясны взаимоотношения между эмоциями [17].

Большинство современных классификаторов используют одношаговую классификацию, которая принимает во внимание сразу все эмоции для классификации. Сходства между близкими эмоциями в эмоциональном пространстве могут быть упущены, а различия между ними могут затемняться.

Для решения этого могут быть использованы многоступенчатые иерархические классификаторы, имеющие структуру двоичного дерева, в котором каждый шаг классификации соответствует одной из размерностей эмоционального пространства, для обеспечения наилучшей возможности разделения речевых признаков на эмоции [18].

Так как эмоции являются достаточно субъективной характеристикой при обработке звукового сигнала и даже люди не всегда могут точно определить преобладающую эмоциональную составляющую в произнесенной фразе, необходимо учитывать неоднозначность принадлежности фразы к определенному эмоциональному состоянию. Эмоция, содержащаяся в звуковом фрагменте, не может всегда быть отнесена только к одному типу, однако она может быть представлена как промежуточная между некоторыми типами эмоций, либо как комбинация нескольких типов эмоций, особенно лежащих близко в пространстве эмоций.

2.3.1 Метод ближайших соседей

Метод ближайших соседей – алгоритм классификации, основанный на оценке сходства объектов. Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки.

Несмотря на простоту, метод ближайших соседей применяется во многих задачах классификации и регрессии. Он неявно опираются на предположение, называемое гипотезой компактности: если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных. В этом случае граница между классами имеет достаточно простую форму, а классы образуют компактно локализованные области в пространстве объектов.

Для повышения надёжности классификации объект относится к тому классу, которому принадлежит большинство из его соседей — ближайших к нему объектов обучающей выборки. В задачах с двумя классами число соседей берут нечётным, чтобы не возникало ситуаций неоднозначности, когда одинаковое число соседей принадлежат разным классам. В задачах с большим числом классов нечётность уже не помогает, и ситуации неоднозначности всё равно могут возникать. Тогда каждому соседу приписывается вес, как правило, убывающий с ростом ранга соседа. Объект относится к тому классу, который набирает больший суммарный вес среди ближайших соседей.

Основные проблемы метода ближайших соседей:

1) выбор числа соседей – при малом количестве соседей метод неустойчив к шумовым выбросам, а при большом количестве чрезмерно устойчив и вырождается в константу. На практике число соседей выбирают методом скользящего контроля;
2) отсев шумов (выбросов) – наличие шумовых выбросов значительно ухудшает качество классификации, а их удаление является непростой задачей;
3) сверхбольшие выборки – метод основан на явном хранении всех обучающих объектов, поэтому необходимо не только хранить большой объем данных, но и уметь быстро находить среди них ближайших соседей. Для этого используются специальные индексы или эффективные структуры данных;
4) выбор метрики – в реальных задачах редко заранее известна лучшая функция расстояния. При этом все признаки должны быть измерены в одном масштабе (нормированы), чтобы признаки с наибольшими числовыми значениями не доминировали в метрике.

Метод ближайших соседей может быть легко интерпретирован и дает хороший результат при применении в системах автоматического распознавания речи.

2.3.2 Метод опорных векторов

Метод опорных векторов – простой и эффективный алгоритм, который имеет хорошую эффективность в сравнении с другими классификаторами. Он является популярным методом обучения для решения задач классификации и регрессии. Метод опорных векторов имеет преимущества при работе с обучающей выборкой небольшого размера. Однако не хватает руководящих принципов по выбору ядра и оптимизации параметров метода. Также недостатком является медленное обучение в случае задачи распознавания многих классов.

Алгоритм заключается в переводе изначальных векторов признаков в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся на обеих сторонах гиперплоскости, отделяющей классы. Отделяющей гиперплоскостью будет гиперплоскость, которая максимизирует расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Основными преимуществами данного метода является то, что это наиболее быстрый метод нахождения решающих функций, который сводится к решению задачи квадратичного программирования в выпуклой области, которая всегда имеет единственное решение. Метод находит разделяющую полосу максимальной ширины, что позволяет в дальнейшем осуществлять более уверенную классификацию.

К недостаткам метода можно отнести чувствительность к шумам и стандартизации данных, а также проблему выбора ядра в случае линейной неразделимости классов.

2.3.3 Модель гауссовых смесей

Модель гауссовых смесей широко используется в решении задачи распознавания дикторов.

Часто в системах, использующих эту модель, используется диагональная матрица ковариации. Также возможно использование одной матрицы ковариации для всех компонентов модели диктора или одной матрицы для всех моделей.

Существует ряд причин для использования моделей гауссовых смесей для решения задачи распознавания эмоций из речи. Одной из них является интуитивное предположение о том, что отдельные компоненты модели могут моделировать некоторое множество акустических признаков.

Другой причиной использования моделей гауссовых смесей является эмпирическое наблюдение, что линейная комбинация гауссовых признаков может представлять большое число классов акустических признаков. Одна из сильных сторон модели гауссовой смеси та, что эти модели могут очень точно аппроксимировать произвольные распределения.

Недостатком использования модели гауссовых смесей является трудность извлечения вектора признаков из каждого кадра, а также анализ выходных данных, так как трудно разделить их на классы.

2.3.4 Скрытые марковские модели

Учет глобальных статистических характеристик сигнала имеет ряд недостатков. В первую очередь, необходимо учитывать то, что они игнорируют изменяющуюся структуру речи, будучи чувствительными к свойствам, которые в данном случае определяются лингвистически. Фонетическое содержание фраз и их структура имеет такое же сильное влияние на выделяемые признаки, как и эмоции, а в некоторых случаях даже большее. Например, вопросительное предложение обычно подразумевает более широкое изменения высоты звука,

чем утвердительное предложение, поэтому для него стандартное отклонение высоты обычно больше. Однако это не имеет никакого отношения к эмоциям, а определяется только структурой предложения. Другим ограничением для использования глобальных статистических параметров является то, что обработка может быть произведена только один раз после того, как фраза будет произнесена целиком. Этот факт ограничивает возможности построения системы автоматического распознавания эмоций, работающей в реальном времени. Также это является основным недостатком, когда эмоциональное состояние изменяется на протяжении произносимой фразы [9].

Различные применения глобальных статистических параметров учитывают то, что этот способ моделирования является лишь отражением кратковременного поведения. Например, вместо использования математических ожиданий и стандартных отклонений кратковременных значений высоты звука или энергии можно исследовать непосредственно функцию плотности распределения вероятности. Плотность распределения вероятности содержит ту же информацию, что и те рассчитанные значения, однако в ней есть гораздо более полная информация обо всем наборе значений признаков.

Если рассматривать моделирование распределения плотности вероятности с помощью модели гауссовых смесей, проблема эквивалентна использованию скрытых марковских моделей с одним состоянием. Скрытые марковские модели имеют долгую историю применения к задаче распознавания речи. Идея, лежащая в основе данной модели, утверждает, что статистические параметры голоса не стационарны во времени. Вместо этого, голос моделируется как последовательность состояний, каждое из которых моделирует различные звуки или звуковые комбинации и имеет свои статистические особенности. Существует два главных преимущества скрытых марковских моделей перед использованием глобальных статистических параметров. Первое из них то, что структура скрытых марковских моделей может быть полезна для обнаружения изменяющегося поведения речи. Второе преимущество – то, что модель скрытых марковских цепей долгое время изучалась в целях использования в задачах распознавания и имеет хорошо оптимизированный и разработанный каркас, включающий в себя алгоритм Баума-Велша и другие [9].

Метод с использованием скрытых марковских моделей является одним из наиболее эффективных методов обработки (распознавания) речевых сигналов является. Скрытая марковская модель – статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами. Главной задачей скрытых марковских моделей является определение неизвестных параметров на основе наблюдаемых. Полученные

параметры могут быть использованы в дальнейшем анализе, например, для распознавания образов.

Применение скрытых марковских моделей в распознавании основывается на следующих предположениях:

1) речевой сигнал может быть сегментирован на фрагменты (состояния), внутри которых сигнал может рассматриваться как стационарный. Переход между этими состояниями осуществляется мгновенно;
2) вероятность появления символа, порождаемого моделью, зависит только от текущего состояния модели и не зависит от предыдущих порожденных символов.

Существует несколько типов скрытых марковских моделей, различающихся по своей топологии.

2.3.5 Нейросетевые методы классификации

Одним из наиболее эффективных методов распознавания речевых сигналов является метод с использованием нейронных сетей, которые состоят из нейронов и с организованными между ними связями. Нейрон является ячейкой нейронной сети. Нейроны могут иметь различные связями между собой: синапсы – однонаправленные входные связи, аксоны – выходные связи нейрона, по которым сигналы поступают на синапсы последующих нейронов. На рисунке 2.1 представлен общий вид нейрона.

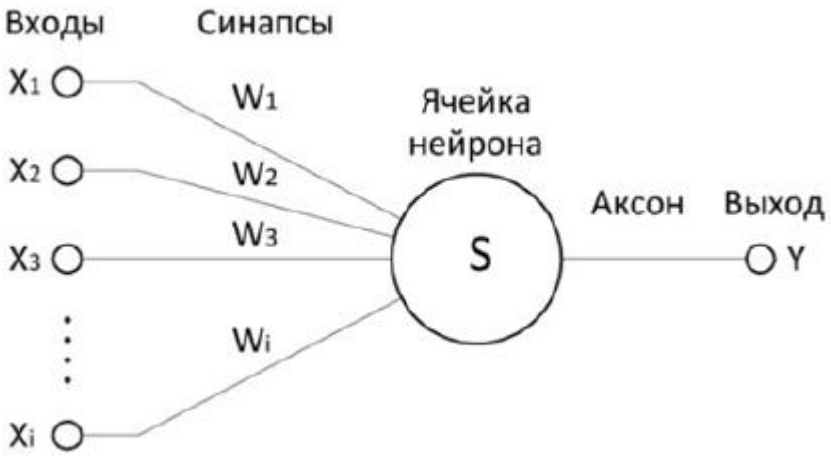

Рисунок 2.1 – Общий вид нейрона

Каждая однонаправленная связь характеризуется весом. Сумма всех входов определяет текущее состояние нейрона.

При использовании нейронных сетей для решения задачи распознавания речевых сигналах необходимо построить соответствующую подходящую для данной задачи сеть, далее обучить ее множеству речевых сигналов – подобрать весовые коэффициенты синапсов для достижения минимизации величины ошибки.

Все признаки, выделенные из фрагментов речи, передаются на вход искусственной нейронной сети, которая состоит из входной матрицы и выходной матрицы, которая отображает эмоциональное состояние каждого предложения, составляющего вход нейронной сети. Искусственные нейронные сети обучаются на обучающей выборке и осуществляют классификацию контрольной выборки, а значение ошибки показывает качество произведенной классификации.

Общий принцип работы искусственной нейронной сети показан на рисунке 2.2. Входные и целевые данные загружаются в нейронную сеть. Входные данные в этом случае представляют собой матрицу признаков, извлеченных из речи. Целевые данные отображают эмоциональное состояние, соответствующее входным данным. Далее входные данные разделяются на категории, называемые обучающей выборкой и контрольной выборкой. При подаче на вход обучающей выборки, параметры классификатора изменяются для обеспечения оптимального значения весовых коэффициентов каждого признака [15].

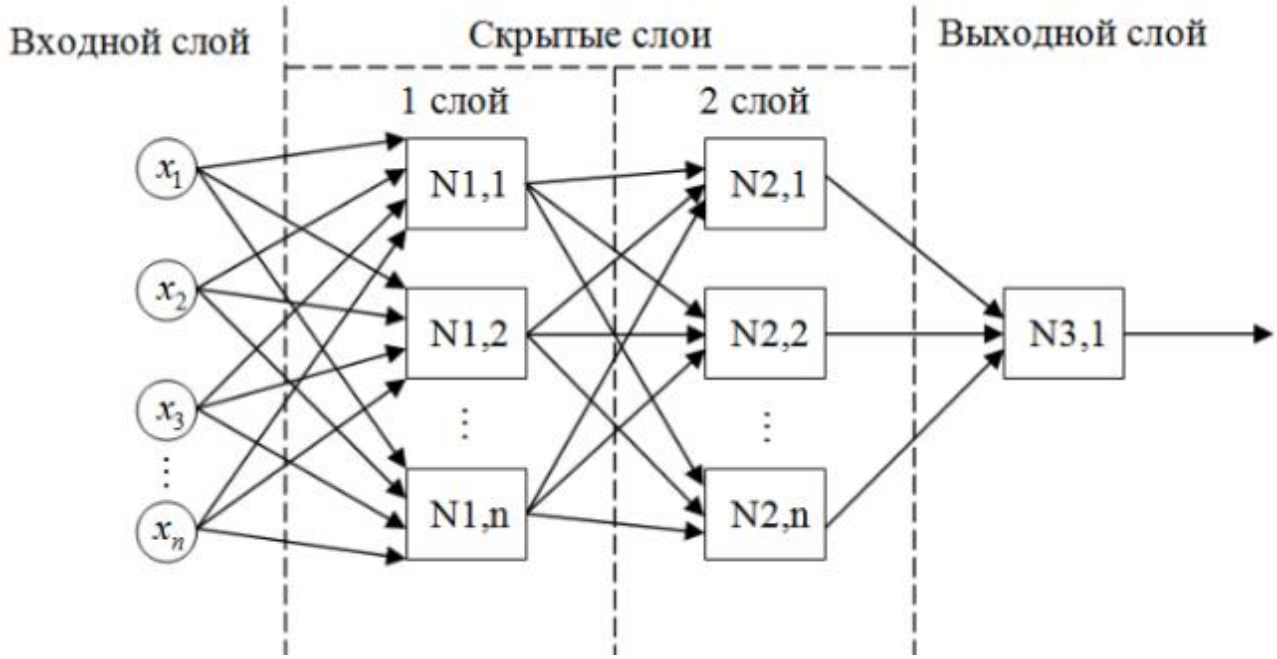

Рисунок 2.2 – Общий принцип работы искусственной нейронной сети

После этого на контрольной выборке тестируется полученная модель и оценивается ошибка классификации. Чем больше количество скрытых слоев в

нейронной сети, тем более сложной получается система и более точным становится полученный результат.

Глубокое обучение является одним из популярных направлений в развитии искусственных нейронных сетей в последние годы. Многообещающей характеристикой глубоких нейронных сетей является то, что они могут выделить высокоуровневые инвариантные признаки из исходных данных, которые могут быть потенциально полезными для распознавания эмоций.

Рекуррентные нейронные сети — вид нейронных сетей, где связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки. В отличие от многослойных персептронов, рекуррентные сети могут использовать свою внутреннюю память для обработки последовательностей произвольной длины. Поэтому рекуррентные нейронные сети применимы в таких задачах, где нечто целостное разбито на сегменты, например: распознавание рукописного текста или распознавание речи.

Наибольшее распространение в задачах распознавания речи получили сети с долговременной и кратковременной памятью (LSTM-сеть). Данная сеть является разновидностью архитектуры рекуррентных нейронных сетей. В отличие от традиционных нейронных сетей, она хорошо приспособлена к обучению на задачах классификации, где важные события разделены интервалами с неопределенной продолжительностью и временными границами. Невосприимчивость к временным разрывам дает данным сетям преимущество по сравнению с другими методами классификации.

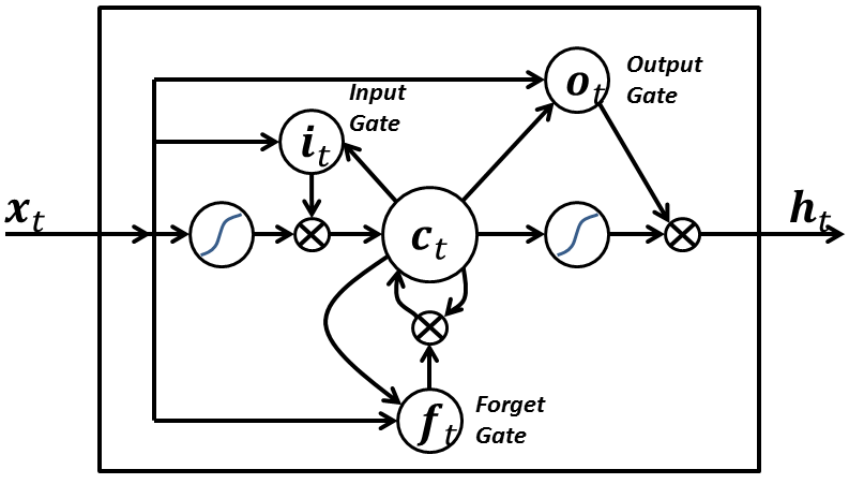

Рисунок 2.3 – Простой LSTM-блок

LSTM-сети состоят из LSTM-модулей, которые способны запоминать значения на как на короткие, так и на длинные промежутки времени. Для минимизации ошибки используется итеративный градиентный спуск. Также

для тренировки могут использоваться комбинация эволюционных алгоритмов для весов в скрытых слоях и метода опорных векторов для весов в выходном слое.

Сверточные нейронные сети – еще один вариант архитектуры нейронных сетей, который может быть использован для решения задачи автоматического распознавания эмоций из речи. Идея, лежащая в основе архитектуры сверточных нейронных сетей, заключается в чередовании сверточных и субдискретизирующих слоев (слоев предвыборки). Структура сети является однонаправленной, то есть не содержит обратных связей, и многослойная.

Работа сверточной нейронной сети интерпретируется как переход от конкретных особенностей к более абстрактным деталям, от которых переходят к абстрактным деталям более высокого уровня, вплоть до выделения необходимых понятий. При этом сеть отбрасывает малосущественные детали и выделяет только необходимые особенности в процессе самонастройки. Однако интерпретация полученных признаков носит скорее иллюстративный характер, так как часто содержание данных признаков малопонятно и трудно.

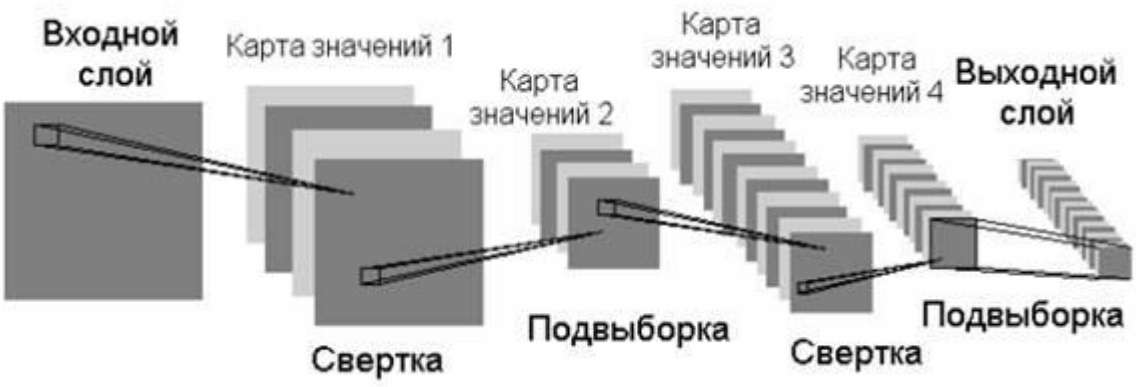

Рисунок 2.4 – Сверточная нейронная сеть

Чаще всего для обучения сверточных нейронных сетей используется метод обратного распространения ошибки и его модификации. Для повышения устойчивости сети и улучшения ее работы, а также для предотвращения переобучения используется метод тренировки с выбрасыванием одиночных нейронов.

В работе [4] предлагается использовать двухэтапный подход, основным отличием которого от предыдущих является то, что нейронная сеть используется не в качестве классификатора, а в качестве генератора признаков. Высказывание, которое необходимо классифицировать, разделяется на пересекающиеся фрагменты, называемые кадрами. Для каждого кадра рассчитываются акустические признаки, по которым находятся вероятности нахождения в одном из эмоциональных состояний. Затем рассчитываются

некоторые статистические параметры от временного ряда вероятностей, по которым с помощью специальной нейронной сети выносится решение об эмоциональном состоянии целого высказывания.

Одним из недостатков этого подхода является то, что не совсем верно предположение о том, что эмоциональное состояние кадра совпадает с эмоциональным состоянием всего высказывания. Однако сложно учесть влияние положения кадра в высказывании на его эмоциональное состояние, так как оно будет достаточно сильно различаться в зависимости от истинной эмоции, длительности высказывания, языка и особенностей речи говорящего. Данную проблему можно решить использованием двухэтапного подхода: сначала используется рекуррентная нейронная сеть для генерации признаков более высокого уровня, таким образом, учитывая временную динамику, а при оценке эмоции всего высказывания используются только наиболее громкие фрагменты. Также на каждой эпохе обучения нейронной сети выполняется корректировка обучающей выборки кадров при помощи скрытой марковской модели, что позволяет частично учесть тот факт, что эмоция целого высказывания может не совпадать с эмоциональной окраской ее конкретного короткого фрагмента.

В большинстве других исследований используется одноэтапный подход. Его идея состоит в подсчете акустических признаков для всего высказывания, а затем обучения этим признакам какой-либо классификационной модели.

2.3.6 Анализ с использованием динамического трансформирования времени

Известно, что речевой сигнал быстро изменяется во времени. Различные произношения одного и того же слова обычно имеют разную длительность, а произношения одного и того же слова одинаковой длительности отличаются в середине из-за различных частей слова, произносимых с разной скоростью. Чтобы получить оценку расхождения между двумя речевыми сигналами, представленными как векторы, должно быть выполнено выравнивание по времени, которое можно реализовать с помощью динамического трансформирования времени.

Данный метод является методом эластичного сравнения вектора наблюдений с хранящимся шаблоном. Вектор наблюдений и шаблон лежат на соответствующих осях сетки (рис. 3). Для каждой ячейки сетки высчитывается разность между соответствующими фрагментами вектора наблюдений и шаблона. Оптимальное выравнивание между вектором наблюдений и шаблоном показано маршрутом, проходящим по сетке.

Метод работает с фрагментами, то есть анализ признаков состоит из обработки вектора признаков в регулярных интервалах. Так как вектор признаков может иметь множество фрагментов, требуются средства расчета локальной оценки расстояния. Оценка расстояния между двумя векторами признаков рассчитывается с помощью евклидового расстояния.

2.4 Оценка результатов обучения

2.4.1 Переобучение

Переобучение — нежелательное явление, возникающее при решении задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке. Переобучение возникает при использовании избыточно сложных моделей.

Эмпирическим риском называется средняя ошибка алгоритма на обучающей выборке. Метод минимизации эмпирического риска наиболее часто применяется для построения алгоритмов обучения. Он состоит в том, чтобы в рамках заданной модели выбрать алгоритм, имеющий минимальное значение средней ошибки на заданной обучающей выборке.

Минимизация эмпирического риска не гарантирует, что вероятность ошибки на тестовых данных будет мала. Для этого можно построить контрпример, который минимизирует эмпирический риск до нуля, но при этом абсолютно не способен обучаться. Получив обучающую выборку, алгоритм запоминает её и строит функцию, которая сравнивает предъявляемый объект с запомненными обучающими объектами. Если предъявляемый объект в точности совпадает с одним из обучающих, то эта функция выдаёт для него запомненный правильный ответ. Иначе выдаётся произвольный ответ. Эмпирический риск алгоритма равен нулю, однако он не восстанавливает зависимость и не обладает никакой способностью к обобщению.

Переобучение появляется именно вследствие минимизации эмпирического риска. Переобучение связано с избыточной сложностью используемой модели. Всегда существует оптимальное значение сложности модели, при котором переобучение минимально.

2.4.2 Скользящий контроль

Для решения проблемы переобучения в задачах машинного обучения используется метод скользящего контроля. При этом фиксируется некоторое множество разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках.

Если выборка независима, то средняя ошибка скользящего контроля даёт несмещённую оценку вероятности ошибки. Это выгодно отличает её от средней ошибки на обучающей выборке, которая может оказаться смещённой (оптимистически заниженной) оценкой вероятности ошибки, что связано с явлением переобучения.

Скользящий контроль является стандартной методикой тестирования и сравнения алгоритмов классификации, регрессии и прогнозирования. Существуют различные варианты скользящего контроля, отличающиеся способами разбиения выборки:

1) полный скользящий контроль;
2) случайные разбиения;
3) контроль на отложенных данных;
4) контроль по отдельным объектам;
5) контроль по блокам.

К недостаткам скользящего контроля относятся большие вычислительные затраты, связанные с необходимостью решения задачи классификации несколько раз. Скользящий контроль не дает информации о том, как строить хорошие алгоритмы обучения.

ГЛАВА 3

РЕАЛИЗАЦИЯ СИСТЕМЫ РАСПОЗНАВАНИЯ ЭМОЦИЙ

3.1 Используемые инструменты

Для реализации системы автоматического распознавания речи был выбран язык Python. Выбор данного языка был обусловлен простотой для изучения и наличием большого количества библиотек для машинного обучения.

Основными библиотеками, использованными для разработки, являются библиотека NumPy (научные вычисления), matplotlib (создание графики), scikit-learn (машинное обучение) и python_speech_festures (вычисление мел-частотных кепстральных коэффициентов) [14].

Библиотека scikit-learn представляет собой реализацию ряда алгоритмов для обучения с учителем и без учителя через интерфейс языка программирования Python. Библиотека распространяется под лицензией «Simplified BSD License» и имеет широкое академическое и коммерческое использование. Для ее использования необходима предварительная установка SciPy (Scientific Python). Библиотека ориентирована в первую очередь на моделирование данных и включает в себя функциональность, которая позволяет решать задачи кластеризации, скользящего контроля, выделения признаков, отбора признаков, снижения размерности, оптимизации параметров алгоритма, множественного обучения, задачи обучения с учителем и др. Для библиотеки имеется обширная документация с примерами кода для различных алгоритмов [14].

Библиотека python_speech_features предназначена для получения повсеместно используемых признаков для систем распознавания речи, таких как мел-частотные кепстральные коэффициенты и др. При необходимости имеется возможность изменения параметров, с которыми вычисляются данные признаки.

3.2 Описание системы

Для реализации системы автоматического распознавания эмоций в речи в качестве источника данных с записями эмоциональной речи была использована

Берлинская база данных эмоциональной речи. База содержит записи высказываний в формате .wav с частотой дискретизации 16 кГц.

Для каждой из записей были выделены признаки. Для выделения признаков были рассчитаны 13 первых мел-частотных кепстральных коэффициента с шириной окна 25 мс и шагом 10 мс. Количество фильтров в гребенке было принято равным 26, а количество точек в быстром преобразовании Фурье 512.

Для полученных мел-частотных кепстральных коэффициентов были рассчитаны значения статистических параметров, такие как минимальное и максимальное значение, математическое ожидание, среднеквадратичное отклонение, медиана, коэффициент асимметрии и эксцесс. Таким образом, размерность вектора признаков составила 91.

Каждый из полученных признаков был стандартизирован, то есть все их значения были преобразованы таким образом, чтобы каждый признак имел математическое ожидание 0 и дисперсию 1. Стандартизация позволяет учитывать различные признаки, выделяемые из речи, в равной степени и исключить более сильное влияние на результат классификации признаков с большими по величине абсолютными значениями [14].

Рисунок 3.1 показывает изменение на протяжении фрагмента записей первых трех мел-частотных кепстральных коэффициентов для записи речи мужчины, произносящего одну и ту же фразу «Das will sie am Mittwoch abgeben», что означает «Она хочет отдать это в среду» в разных эмоциональных состояниях (счастья и грусти).

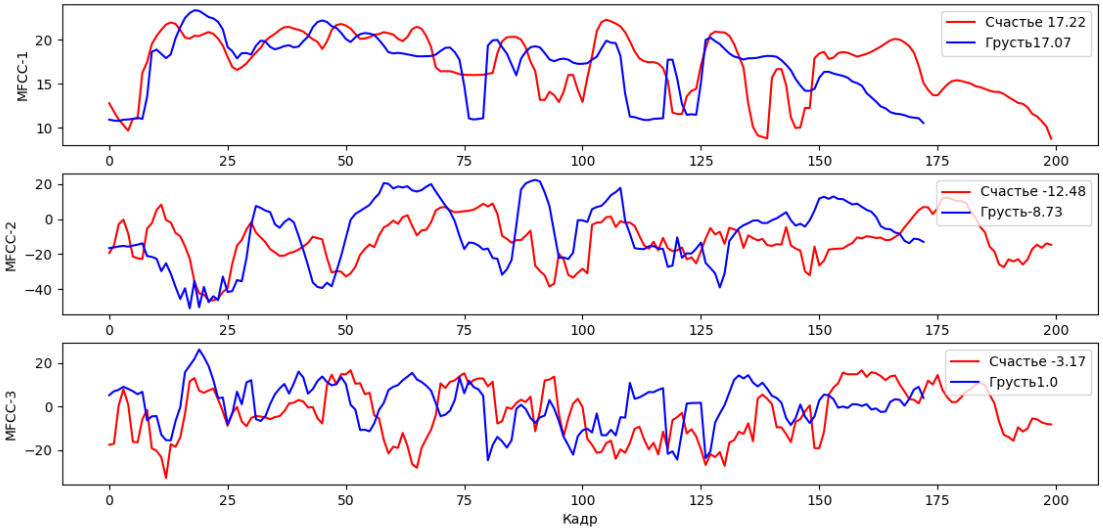

Рисунок 3.1 – Изменение первых трех мел-частотных кепстральных коэффициентов для двух эмоциональных состояний

Рядом с обозначением эмоции указано значения математического ожидания соответствующего коэффициента. Из данного рисунка видно, что математическое ожидание первого коэффициента выше, когда фраза произносится в состоянии счастья, и ниже для второго и третьего коэффициентов. Для того чтобы более детально проанализировать это различие, для каждого из мел-частотных кепстральных коэффициентов рассчитывается ряд различных статистических характеристик.

В качестве алгоритмов классификации были выбраны метод ближайших соседей и метод опорных векторов, как одни из наиболее простых алгоритмов, дающих, тем не менее, приемлемый результат в решении задачи классификации отрывков эмоциональной речи.

Так как размерность вектора признаков равна 91, то каждую из записей эмоциональной речи из размеченной базы данных можно представить точкой в пространстве такой же размерности. Каждой базовой эмоции в данном пространстве соответствует определенная область. Нахождение границ между данными областями соответствует решению задачи классификации записей.

Так как пространство данной размерности сложно визуализировать, для наглядности представим все записи в двумерном пространстве признаков. В качестве признаков выбраны математические ожидания первого и второго мел-частотного кепстрального коэффициента. Результат визуализации для двух эмоциональных состояний (злость и скука) представлен на рисунке 3.2. Даже исходя из рассмотрения записей эмоциональной речи в двумерном пространстве признаков видно, что различные эмоциональные состояния группируются вокруг различных областей на плоскости.

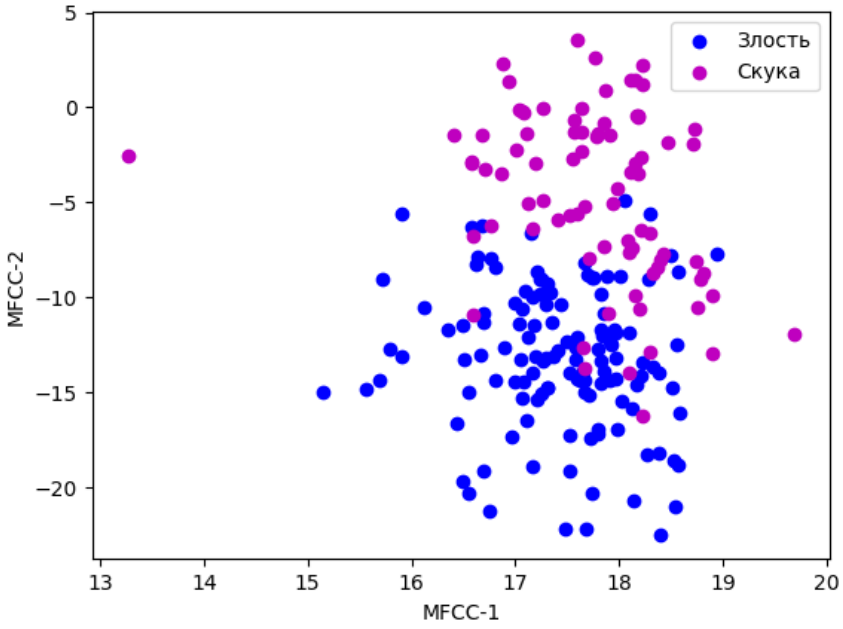

Рисунок 3.2 – Визуализация записей с соответствующими эмоциями в пространстве двух признаков для двух эмоциональных состояний

3.3 Результаты распознавания

Для метода ближайших соседей применялся полный скользящий контроль по методу LOO (leave-one-out), что означает контроль по отдельным объектам. При этом каждый объект ровно один раз участвует в контроле, а длина обучающей выборки только на единицу меньше полной выборки [14]. Для нахождения расстояния между объектами использовалась евклидова метрика.

В результате моделирования была установлена зависимость точности распознавания от количества ближайших соседей. На рисунке 3.2 показана зависимость оценки скользящего контроля для классификации от числа соседей. Как видно из рисунка, наилучший результат был получен при значении количества ближайших соседей 10. Оценка скользящего контроля для распознавания эмоций при этом составила 69,3%.

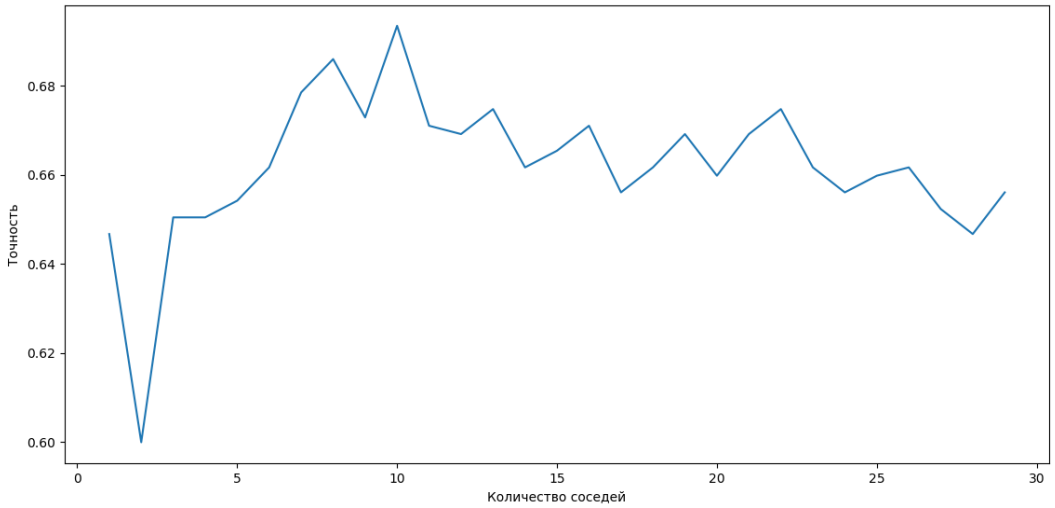

Рисунок 3.3 – Точность классификации в зависимости от k для метода ближайших соседей

Матрица неточности для классификатора с числом соседей, равным 10, полученная по результатам скользящего контроля, приведена в таблице 3.1. Также в таблице дополнительно приведена точность и полнота для каждого класса эмоций, а также итоговая точность классификации.

Из анализа полученных результатов можно увидеть, что чаще всего встречаются ошибки классификации между классами злость и счастье, а также между скукой и нейтральной эмоцией. Также эмоции страха и отвращение часто принимаются за другие эмоциональные состояния.

Таблица 3.1 – Результат классификации методом ближайших соседей при k=10

		Распознанный класс							Полнота, %
		A	B	D	F	H	S	N	
Действительный класс	Злость (A)	113	0	1	0	13	0	0	89,0
	Скука (B)	4	37	1	1	0	7	31	45,7
	Отвращение (D)	5	1	31	3	3	1	2	67,4
	Страх (F)	6	4	1	46	6	3	3	66,7
	Счастье (H)	32	0	3	2	34	0	0	47,9
	Грусть (S)	0	4	0	0	0	55	3	88,7
	Нейтральная (N)	2	14	2	3	2	1	55	69,6
Точность, %		69,8	61,7	79,5	83,6	58,6	82,1	58,5	69,3

Также в работе сравнивалась точность классификации с применением метода ближайших соседей и метода опорных векторов с линейным ядром. Для метода ближайших соседей подбиралось оптимальное значения количества соседей, которое изменялось от 1 до 30, а для метода опорных векторов – параметр C, описывающий ошибку классификации, который изменялся от 10^{-6} до 10^{10} в логарифмическом масштабе. При проверке скользящим контролем лучшими классификаторами оказались метод ближайших соседей с k=29 (точность составила 55,33%) и метод опорных векторов с C=0,0278 (точность – 57,57%).

Оценка полного скользящего контроля для классификатора, использующего метод опорных векторов, по методу LOO составила 72,3%. Таким образом, метод опорных векторов оказался эффективнее метода ближайших соседей при решении задачи классификации эмоций в речи.

Матрицы неточностей для классификатора, использующего метод опорных векторов, приведена в таблице 3.2.

Таблица 3.2 – Результат классификации методом опорных векторов при C=0,0278

		Распознанный класс							Полнота, %
		A	B	D	F	H	S	N	
Действительный класс	Злость (A)	108	0	2	1	16	0	0	85,0
	Скука (B)	1	51	2	0	1	4	22	63,0
	Отвращение (D)	4	1	29	5	3	0	4	63,0
	Страх (F)	4	4	4	48	5	3	1	69,6
	Счастье (H)	19	1	2	4	43	1	1	60,6
	Грусть (S)	0	7	1	1	0	52	1	83,9
	Нейтральная (N)	1	19	1	1	0	1	56	70,9
Точность, %		78,8	61,4	70,7	80,0	63,2	85,2	65,9	72,3

Для полученного классификатора была оценена информативность используемых признаков. Для этого используются алгоритмы, основанные на переборе подмножеств признаков с целью нахождения подмножества, на которых обученная модель дает наилучший результат. Одним из таких алгоритмов является алгоритм Recursive Feature Elimination. Данный алгоритм выбирает признаки рекурсивно, рассматривая все меньшие наборы признаков на каждом шаге. На каждом шаге наименее важные признаки выбрасываются из текущего набора признаков. Алгоритм выполняется, пока количество признаков не достигнет желаемого [14].

Для полученного классификатора, использующего метод опорных векторов, была получена зависимость точности распознавания обучающей выборки от количества признаков, отобранных из исходного набора процедурой рекурсивного удаления признаков. Данная зависимость приведена на рисунке 3.4.

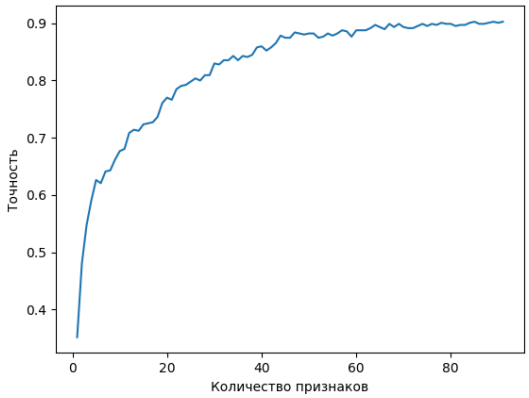

Рисунок 3.4 – Зависимость точности от количества признаков для метода опорных векторов

Из графика видно, что точность распознавания практически не изменяется при увеличении количества признаков более 60. Это показывает, что последующие признаки являются малоинформативными для распознавания эмоций. Список 10 наиболее информативных признаков приведен в таблице 3.3.

Таблица 3.3 – Наиболее информативные признаки для метода опорных векторов

№	Статистический параметр	Номер коэффициента
1	Математическое ожидание	2
2		3
3		5
4		9
5	Среднеквадратичное отклонение	3
6	Минимум	3
7		6
8	Медиана	2
9		4
10		6

ЗАКЛЮЧЕНИЕ

Разработка системы автоматического распознавания эмоций из речи является относительно новой областью исследований и является сложной задачей классификации в машинном обучении из-за ряда особенностей, присущих ей.

Для решения данной задачи была разработана система, состоящая из ряда этапов, таких как использование базы данных эмоциональной речи, выделение признаков из записанных образцов, отбор значимых признаков и их нормализация, обучение различных классификаторов на основе тестовых данных и подбор оптимальных параметров классификаторов.

При решении задачи было проанализировано различие между величинами статистических характеристик значений мел-частотных кепстральных коэффициентов для записей речи с различными эмоциональными состояниями.

В ходе работы были определены признаки и алгоритмы классификации, которые могут быть эффективно использованы при моделировании систем автоматического распознавания эмоций в речи, а также определены оптимальные параметры для данных алгоритмов.

Данная система может иметь большое количество применений в различных областях человеческой деятельности, таких как системы выявления негативных эмоциональных состояний, синтез эмоциональной речи, определение степени удовлетворенности клиентов услугами.

Развитие методов машинного обучения и алгоритмов обработки речи позволило добиться высокого качества распознавания эмоций.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. А.К.Амурадов, П.П.Чураков, «Обзор и классификация методов обработки речевых сигналов в системах распознавания речи». Измерение. Мониторинг. Управление. Контроль, 2015

2. R.Banse, K.R.Scherer, "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology, Vol.70, 614-636, 1996

3. T.Banziger, K.R.Scherer, "The role of intonation in emotional expression", Speech Communication, Vol.46, 252-267, 2005

4. P. Ekman, W. Friesen, “Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues”, Trans. Piter Publishing house, Russia, 2010.

5. S.Kim, P.Georgiou, S.Lee, S.Narayanan. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece, 2007

6. J.Lee, I.Tashev. “High Level feature representation using recurrent neural network for speech emotion recognition”, Sixteenth Annual Conference of the International Speech Communication Association, 2015

7. M.Mantyla, B.Adams, G.Destefanis, D.Graxiotin, M.Ortu. “Mining Valence, Arousal, and Dominance - Possibilities for Detecting Burnout and Productivity”, Proceedings of the I3th International Workshop on mining Software Repositories, 247-258, 2016

8. H. McGurk, J. MacDonald. “Hearing lips and seeing noices”, Nature, Vol.264(5588), 746-748

9. A.Nougueiras, A.Moreno, A.Bonafonte, J.B.Marino. “Speech Emotion Recognition Using Hidden Markov Models”, Eurospeech 2001 – Scandinavia, 2001

10. V.A Petrushin, "Emotional Recognition in Speech Signal: Experimental Study, Development, and Application", ICSLP-2000, Vol.2, 222-225, 2000

11. R.Plutchik, “The Nature of Emotions”. American Scientist, 2011.

12. L.R.Rabiner and R.W.Schafer. "Digital processing of speech signals", Englewood Cliffs; London: Prentice-Hall, 1978

13. L.R.Rabiner and B.H.Juang. "Fundamentals of Speech Recognition", Upper Saddle River; NJ: Prentice-Hall, 1993

14. W.Richert, L.P.Coelho, "Building Machine Learning Systems With Python", Packt Publishing, 2013

15. A.Shaw, R.K.Vardhan, S.Saxena. “Emotion Recognition and Classification in Speech using Artificial Neural Networks”, International Journal of Computer Applications, Vol.145, 4-9, 2016

16. D.Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", Speech Coding & Synthesis, 1995

17. S.Vaishnav, S.Mitra. “Speech Emotion Recognition: A Rewiew”, IRJET, Vol.03, 313-316, 2016

18.Z.Xiao. “Recognition of Emotions in. Audio Signals”. Doctoral dissertation. L’ecole Centrale de Lyon, 2008

19. F.Yu, E.Chang, Y.Xu, H.Shum, "Emotion detection from speech to enrich multimedia content", Lecture Notes In Computer Science, Vol.2195, 550-557, 2001

ПРИЛОЖЕНИЕ А

Исходный код разработанной программы

load_data.py

```
from python_speech_features import mfcc
from scipy.stats import kurtosis
from scipy.stats import skew
import scipy.io.wavfile as wav
import numpy as np
import os
import csv
import re

emotions_map = {'W': 1, 'L': 2, 'E': 3, 'A': 4, 'F': 5, 'T': 6, 'N': 7}
folder_path = 'D:/docs/mfcc/download/wav'
output_file = 'D:/docs/features.csv'


def read_file(folder, filename):
    file_path = os.path.normpath(os.path.join(folder, filename))
    (rate, sig) = wav.read(file_path)
    return rate, sig


def extract_features(rate, sig):
    mfcc_coeffs = mfcc(sig, rate)
    mfcc_features = calculate_features(mfcc_coeffs)
    return mfcc_features


def calculate_features(mfcc_coeffs):
    mfcc_mean = np.mean(mfcc_coeffs, axis=0).tolist()
    mfcc_std = np.std(mfcc_coeffs, axis=0).tolist()
    mfcc_min = np.min(mfcc_coeffs, axis=0).tolist()
    mfcc_max = np.max(mfcc_coeffs, axis=0).tolist()
    mfcc_median = np.median(mfcc_coeffs, axis=0).tolist()
    mfcc_skew = skew(mfcc_coeffs).tolist()
    mfcc_kurtosis = kurtosis(mfcc_coeffs).tolist()
    features = mfcc_mean + mfcc_std + mfcc_min + mfcc_max + mfcc_median +
mfcc_skew + mfcc_kurtosis
    return features


def save_to_file(features, file_path):
    out_file = open(os.path.normpath(file_path), 'w')
    writer = csv.writer(out_file)
    for row in features:
        writer.writerow(row)
    out_file.close()


def load_features():
    folder = os.path.normpath(folder_path)
    mfcc_features = []
    for filename in os.listdir(folder):
        emotion = re.findall(r'\w{5}(\w)\w\.wav', filename)
        if emotion:
            (rate, sig) = read_file(folder, filename)
            features = extract_features(rate, sig)
            features.append(emotions_map[emotion[0]])
```

```
            mfcc_features.append(features)
    save_to_file(mfcc_features, output_file)
    return mfcc_features
```

classify.py

```
from sklearn import metrics, preprocessing
from sklearn.model_selection import KFold, cross_val_score, GridSearchCV
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
import matplotlib.pyplot as plt
import load_data
import numpy as np

def scale_data(X):
    scaled_X = preprocessing.scale(X);
    return scaled_X


def classify_knn(x, y, k):
    model = KNeighborsClassifier(n_neighbors=k)
    model.fit(x, y)
    print('test')
    return model


def classify_svm(x, y):
    model = SVC()
    model.fit(x, y)
    return model


def predict(model, x, y):
    expected = y
    predicted = model.predict(x)
    print(metrics.classification_report(expected, predicted))
    print(metrics.confusion_matrix(expected, predicted))


def cross_validation(x, y):
    cv_scores = []
    n = len(x)
    for k in range(1, 30):
        kf = KFold(n_splits=n, shuffle=False)
        knn = KNeighborsClassifier(n_neighbors=k)
        scores = cross_val_score(knn, x, y, cv=kf, scoring='accuracy')
        tmp = scores.mean()
        cv_scores.append(tmp)
        print('{} : {}'.format(k, tmp))
    return cv_scores


def cross_validation_results(cv_scores):
    optimal_k = cv_scores.index(max(cv_scores)) + 1
    print("The optimal number of neighbors is %d" % optimal_k)
    plt.plot(cv_scores)
    plt.xlabel('Number of neighbors')
    plt.ylabel('Accuracy')
    plt.show()
    return optimal_k


svc = SVC(kernel='linear')
C_s = np.logspace(-6, -1, 10)
clf = GridSearchCV(estimator=svc, param_grid=dict(C=C_s))
```

```
clf.fit(X, y)
print(clf.best_score_)
print(clf.best_estimator_.C)
predicted = clf.predict(X)
print(metrics.confusion_matrix(y, predicted))
print(metrics.classification_report(y, predicted))

knn = KNeighborsClassifier();
k_s = range(1, 30, 1)
clf = GridSearchCV(estimator=knn, param_grid=dict(n_neighbors=k_s))
clf.fit(X, y)
print(clf.best_score_)
print(clf.best_estimator_.n_neighbors)
predicted = clf.predict(X)
print(metrics.confusion_matrix(y, predicted))
print(metrics.classification_report(y, predicted))
```

visualization.py

```
import re
import matplotlib.pyplot as plt
import scipy.io.wavfile as wav
from python_speech_features import mfcc
import os
import numpy as np
from collections import defaultdict

DIR_PATH = '../docs/mfcc/download/wav'
emotions_color = {'W': 'g', 'L': 'c', 'E': 'y', 'A': 'r', 'F': 'b', 'T': 'k',
'N': 'm'}
emotions_label = {'W': 'Злость', 'L': 'Скука', 'E': 'Отвращение', 'A': 'Страх',
'F': 'Счастье', 'T': 'Грусть', 'N': 'Нейтральная'}


def get_data(file):
    file_path = os.path.normpath(os.path.join(DIR_PATH, file))
    (rate, sig) = wav.read(file_path)
    mfcc_coefficients = mfcc(sig, rate, numcep=3)
    return mfcc_coefficients

if __name__ == "__main__":
    mfcc_happiness = get_data("03a02Fc.wav")
    mfcc_sadness = get_data("03a02Ta.wav")
    fig, axes = plt.subplots(nrows=3, ncols=1)
    for i in range(len(axes)):
        mean_happiness = str(round(np.mean(mfcc_happiness[:, i]), 2))
        mean_sadness = str(round(np.mean(mfcc_sadness[:, i]), 2))
        axes[i].plot(mfcc_happiness[:, i], 'r', label = "Счастье " +
mean_happiness)
        axes[i].plot(mfcc_sadness[:, i], 'b', label = "Грусть" + mean_sadness)
        axes[i].set_xlabel("Кадр")
        axes[i].set_ylabel("MFCC-" + str(i + 1))
        axes[i].legend(loc=1)
    plt.show()

    emotions = defaultdict(list)
    plt.figure()
    for filename in os.listdir(DIR_PATH):
        emotion = re.findall(r'\w{5}(\w)\w\.wav', filename)
        if emotion:
            features = get_data(filename)
            mean_first = np.mean(features[:, 0])
            mean_second = np.mean(features[:, 1])
            emotions[emotion[0]].append([mean_first, mean_second])
    for k, v in emotions.items():
```

```
        plt.scatter([row[0] for row in v], [row[1] for row in v],
c=emotions_color[k], label=emotions_label[k])
    plt.xlabel("MFCC-1")
    plt.ylabel("MFCC-2")
    plt.legend(loc=0)
    plt.show()
```