Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика и системы управления»

КАФЕДРА ИУ-7 «Программное обеспечение эвм и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

«Метод распознавания эмоций по звучащей речи на основе скрытой марковской модели»

Студент ИУ7-76Б _____ Т. А. Казаева

(Подпись, дата)

Руководитель _____ Ю. В. Строганов

(Подпись, дата)

2023 г.

РЕФЕРАТ

Расчетно–пояснительная записка 39 с., 6 рис., 4 табл., 32 ист., 0 прил.

СОДЕРЖАНИЕ

РЕФЕРАТ 1

ВВЕДЕНИЕ 5

1 Аналитический раздел 6
1.1. Категоризация эмоциональных данных 6
1.1.1. Дискретное пространство эмоций 6
1.1.2. Многомерное пространство эмоций 6
1.1.3. Гибридное пространство эмоций 7
1.2. Наборы данных речевых эмоций 8
1.2.1. Наборы данных на иностранных языках 8
1.2.2. Наборы данных на русском языке 8
1.2.3. Интонационный контур при выражении эмоций 9
1.3. Выделение информативных признаков 10
1.3.1. Просодические характеристики 11
1.3.1.1. Частота основного тона 11
1.3.1.2. Инетнсивность, темп речи и паузация 14
1.3.2. Спектральные характеристики 15
1.3.2.1. Мел-кепстральные коэффициенты 15
1.3.2.2. Энергетические и пертурбационные параметры 17
1.3.2.3. Частоты первых формант речевого сигнала . . 18
1.4. Классификаторы, используемые в анализе речи 19
1.4.1. Скрытая марковская модель 20
1.4.2. Искусственная нейронная сеть 21
1.5. Постановка задачи 22

2 Конструкторский раздел **24**
2.1. Общая схема метода 24
2.2. Проектирование ключевых модулей системы 25
2.2.1. Формирование вектора информативных признаков . . . 25
2.2.2. Кластеризация 26
2.2.3. Создание и обучение скрытых марковских моделей . . . 27
2.2.4. Определение эмоции из аудиосигнала 30

2.3. Описание используемого набора данных 30
2.3.1. Разметка и структура набора 30
2.3.2. Содержание набора данных 31
2.4. Проектирование отношений сущностей 31
2.5. Проектирование клиентского приложения 32
2.6. Вывод 32

3 Технологический раздел 33
3.1. Выбор средств программной реализации 33
3.1.1. Выбор языка программирования 33
3.1.2. Выбор СУБД 33
3.1.3. Модули программного обеспечения 33
3.1.4. Модуль получения вектора информативных признаков сигнала 33
3.1.5. Модуль формирования вектора информативных признаков 33
3.1.6. Модуль кластеризации 33
3.1.7. Модуль создания и обучения скрытых марковских моделей 33

4 Исследовательский раздел 34

ЗАКЛЮЧЕНИЕ 35

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 38

ПРИЛОЖЕНИЕ А 39

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящей расчетно-пояснительной записке применяют следующие термины с соответствующими определениями.

1) Фреймы – отрезки аудиосигнала длительности как правило 10-40 мс, идущие «внахлест», то есть таким образом, чтобы начало очередного фрейма пересекалось с концом предыдущего.

2) Паралингвистически признаки речи – невербальные признаки, передающие совместно с вербальными смысловую информацию в составе речевого сообщения.

3) Частота основного тона – частота колебания голосовых связок при произнесении тоновых звуков.

4) Джиттер – мера возмущений частоты основного тона, показывающая непроизвольные изменения в частоте смежных вибрационных циклов голосовых складок.

5) Шиммер – мера аналогичная джиттеру, только характеризующая пертурбации амплитуд сигнала на смежных циклах колебаний основного тона.

6) Форманты – пики в огибающей спектра звука, создаваемые акустическими резонансами в голосовом тракте.

7) Кепстр – преобразование Фурье от логарифма спектра мощности.

8) Мел – единица измерения частоты звука, основанная на статистической обработке большого числа данных о субъективном восприятии высоты звуковых тонов.

9) Редукция (*в лингвистике*) – ощущаемое человеческим ухом изменение звуковых характеристик речевых элементов, вызванное их безударным положением по отношению к ударным элементам.

10) Кластеризация – разделение множества входных векторов данных на кластеры (группы) по степени схожести друг с другом.

ВВЕДЕНИЕ

хорошие пельмени это очень вкусно

1 Аналитический раздел

1.1. Категоризация эмоциональных данных

Одной из главных проблем в исследованиях, связанных с определением эмоционального состояния диктора по голосу, является отсутствие четкого определения эмоции. Подход к классификации эмоций влияет на процесс аннотирования. Сегодня широко используются три подхода к категоризации эмоциональных данных: дискретный, многомерный и гибридный.

1.1.1. Дискретное пространство эмоций

Дискретный подход основан на выделении фундаментальных (базовых) эмоций, сочетания которых порождают разнообразие эмоциональных явлений. Разные авторы называют разное число таких эмоций – от двух до десяти. П. Экман на основе изучения лицевой экспрессии выделяет пять базовых эмоций: гнев, страх, отвращение, печаль и радость. Первоначальная версия 1999 года также включала удивление. [1; 2] Р. Плутчик [3] выделяет восемь базисных эмоций, деля их на четыре пары, каждая из которых связана с определенным действием: страх, уныние, удивление и т. д.

На сегодняшний день существование базовых эмоций ставится под сомнение. Теория встречает ряд концептуальных проблем, таких как, например, эмпирическое определение набора базовых эмоций или критерии синхронизации эмоциональных реакций. Однако, многие решения в области автоматического детектирования эмоций основаны на дискретной модели эмоциональной сферы. Например, решение компании «Affectiva». [4]

1.1.2. Многомерное пространство эмоций

Многомерное пространство представляет собой эмоции в координатном многомерном пространстве. В качестве ее источника рассматривают идею В. Вундта о том, что многогранность чувств человека можно описать с помощью трех измерений: удовольствие-неудовольствие, расслабление-напряжение, возбуждение-успокоение. Вундт заключил, [5] что эти измерения охватывают все разнообразие эмоциональных состояний. Данные для этой теории были получены с помощью метода интроспекции.

Эмоциональная сфера представляется как многомерное пространство, об-

разованное некоторым количеством осей координат. Оси задаются полюсами первичных характеристик эмоций. Отдельные эмоции – это точки, местоположение которых в «эмоциональном» пространстве определяется степенью выраженности этих параметров.

Один из примеров описываемого подхода – модель Дж. Рассела. В ней водится двумерный базис, в котором каждая эмоция характеризуется валентностью (*англ. valence*) и интенсивностью (*англ. arousal*). Измерение валентности отражает то, насколько хорошо человек ощущает себя на уровне субъективного переживания от максимального неудовольствия до максимального удовольствия. Измерение активации связано с субъективным чувством энергии и ранжируется в диапазоне от дремоты до бурного возбуждения. Такой подход используется, например, в наборе данных «RECOLA» [6].

Аналогично вопросу о количестве эмоций в дискретной модели, вопрос о количестве измерений остается открытым. Использование только двух критикуется на том основании, что они не позволяют устанавливать различия между отдельными эмоциональными состояниями (например, страх, гнев, ревность, презрение и др. имеют отрицательную валентность и высокую активацию).

1.1.3. Гибридное пространство эмоций

Гибридная модель представляет собой комбинацию дискретной и многомерной модели. Примером такой модели являются «Песочные часы эмоций», предложенные Камбрией, Ливингстоном и Хуссейном. [7]

Согласно этой классификации, в отдельной области n-мерного эмоционального пространства различия между эмоциями могут определяться в терминах измерений, имеющих отношение к этой области. Эмоции могут быть сопоставимы по измерениям внутри и вне категорий, и каждая категория может иметь свои отличительные признаки. [8] Каждое измерение характеризуется шестью уровнями силы, с которой выражены эмоции. Данные уровни обозначаются набором из двадцати четырех эмоций. Поэтому совершенно любая эмоция может рассматриваться как и фиксированное состояние, так и часть пространства, связанная с другими эмоциями нелинейными отношениями.

1.2. Наборы данных речевых эмоций

1.2.1. Наборы данных на иностранных языках

Набор данных RAVDESS. [9] Набор содержит записи 24 профессиональных актеров (12 мужчин и 12 женщин), озвучивающих две одинаковые фразы на английском языке с североамериканским акцентом в двух вариантах: речь и пение. На каждого актера набор предоставляет 60 записей. Использовано дискретное эмоциональное пространство, состоящее из семи эмоций: спокойствие, гнев, страх, отвращение и радость. Каждая фраза представлена двумя уровнями эмоциональной интенсивности для каждой из эмоций и безэмоционально.

Набор данных SAVEE [10] *(англ. Surrey Audio-Visual Expressed Emotion)* состоит из речи четырех актеров мужского пола, говорящих на английском с британским акцентом. Эмоциональные типы для каждого высказывания соответствуют одной из шести базовых эмоций (радость, печаль, гнев, удивление, страх, отвращение) или нейтральному состоянию. Всего было записано 15 уникальных фраз, каждая фраза записывалась безэмоционально дважды.

Набор данных Emo-DB. [11] Немецкие актеры (5 женщин и 5 мужчин) имитировали эмоции, произнося фразы, интонационно отражающие различные эмоции (гнев, радость, печаль, страх), а также произнести некоторые из них так, чтобы они не несли никакой эмоциональной нагрузки. Смысл фраз не соответствовал интонационному оформлению. Запись файлов для базы данных проводилась в звукоизолированной комнате. Для оценки качества записанной речи в Берлинском университете был проведен перцептивный тест которым было предложено оценить, к какой из эмоций относится прослушанная единожды запись. Сообщения с уровнем распознавания выше 80% и естественностью звучания свыше 60% вошли в итоговый набор.

Набор данных TESS [12] *(англ. Toronto emotional speech set)* содержит 2800 звуковых дорожек формата «WAV». Набор озвучен только женскими голосами и размечен по 6 базовым эмоциям: гнев, отвращение, страх, счастье, печаль, удивление. Также присутствуют записи безэмоциональной речи.

1.2.2. Наборы данных на русском языке

Набор данных RUSLANA. [13] Является первым рускоязычным эмоциональным набором данных. Содержит записи 61 человека (12 мужчин и 49

женщин), которые произносили десять предложений с выражением следующих эмоциональных состояний: удивление, счастье, гнев, грусть и страх. Также предложения были записаны безэмоционально. Таким образом, в сумме база содержит 3 660 записей. Общая продолжительность аудиозаписей составляет более 31 часа.

Набор данных DUSHA [14] – русскоязычный набор эмоциональных данных. Состоит из нарезки русскоязычных подкастов, содержащие до пяти слов. Все аудиозаписи были сделаны на профессиональные микрофоны. Разметка осуществлялась также согласно дискретной модели и содержала следующие эмоции: радость, грусть, злость. Также присутствуют безэмоциональные записи. Набор содержит примерно из 300 000 аудиозаписей, суммарная длительность которых около 350 часов. В разметке содержится 1,5 млн записей.

Русскоязычный эмоциональный корпус (REC) состоит из 295 видеозаписей университетских зачетов и экзаменов и 510 видеозаписей общения с клиентами в службе одного окна ГУ ИС г. Москвы. К настоящему моменту размечено 192 файла. Разметка содержит информацию о мимике и жестах, которая размещена на временной шкале. Учитывается мимика двух органов: глаза (взгляды вверх и по сторонам продолжительностью больше 0,5-1 с) и рот (неречевые движения губ, манипуляции языком и движения челюстью). Движения рук представлены как самостоятельные, так и с использованием сторонних предметов, тела или одежды.

1.2.3. Интонационный контур при выражении эмоций

Звучащие предложения обладают интонацией: повествовательной, вопросительной, ответной, перечислительной, восклицательной и т.п. Предполагается, что существует связь между выражением эмоции и использованием одного из существующих в русском языке интонационного контура. [15] В русском языке выделяют 7 интонационных контуров (ИК): [16]

– **ИК-1** характеризуется понижением тона на ударной части:
«Анна стоит на мосту. Наташа поет.»,
используется для выражения завершенности в повествовательных предложениях;

– при **ИК-2** ударная часть произносится с некоторым повышением тона:

«Кто пьет сок? Как поет Наташа?»,
наблюдается в вопросе с вопросительными словами;

– **ИК-3** характеризуется значительным повышением тона на ударной части:
«Это Антон? Ее зовут Наташа?»,
наблюдается в вопросе без вопросительных слов;

– **ИК-4** характеризуется повышением тона, продолжающееся на безударных слогах:
«А вы? А это?»,
наблюдается в вопросе без вопросительных слов;

– **ИК-5** используется при выражении оценки в предложениях с местоименными словами:
«Какой сегодня день!»,
наблюдается повышение тона на ударной части;

– **ИК-6**, аналогично ИК5, используется при выражении оценки в предложениях с местоименными словами, однако повышение тона происходит на ударной части и продолжается на заударной части:
«Какой сок вкусный!»;

– **ИК-7**, аналогично ИК-1 используется для выражения завершенности в повествовательных предложениях:
«И Антон стоит на мосту.»,
но ударная часть, в отличие от ИК-1, эмоционально окрашена.

При помощи интонационного контура выражаются различные эмоциональные оттенки, которые всегда отражают состояние говорящего или его желание определенным образом воздействовать на слушающего. Таким образом, возможно существование зависимости между выражением определенной эмоции и выбором одного из семи интонационных контуров. Учет интонационного контура в наборе данных, который будет использован классификатором, может способствовать выявлению этой зависимости.

1.3. Выделение информативных признаков

Важной особенностью речевого сигнала является его условная стационарность на небольших промежутках (от 20 до 40 мс). [17] По этой причине для

оцифровки сигнал разделяется на фреймы. Деление происходит таким образом, чтобы каждая точка перекрывалась дважды. [18]

Задача распознавания эмоций решается непосредственно по оцифрованному сигналу. Выделяется два этапа решения задачи: выделение и отбор информативных признаков и классификация (сопоставление признаков).

Даже при работе с фреймами, сигнал содержит много избыточной для анализа информации. Поэтому для того, чтобы привести сигнал в вид, который будет использован алгоритмом распознавания, требуется выделить набор информативных признаков речевого сигнала. К выделяемому набору признаков предъявляются следующие требования: [19]

– с помощью выделенного набора признаков можно получить наиболее значимую информацию из акустического сигнала;

– размер выборки должен быть минимальным для увеличения быстродействия разрабатываемой системы распознавания эмоций.

Характеристики речевого сигнала, использующиеся для определения эмоций по речи, можно разделить на две группы: просодические и спекртальные.

Просодическими признаками, содержащими информацию об эмоции, являются характеристики, основанные на количественной оценке ЧОТ, интенсивность (энергия) речевого сигнала, темп речи и паузация. К спектральным признакам можно отнести различные кепстральные (например, мел-кепстральные) коэффициенты, частоты первых формант речевого сигнала и их среднеквадратические отклонения, энергетические (джиттер и шиммер) и пертурбационные параметры.

1.3.1. Просодические характеристики

1.3.1.1. Частота основного тона

Значение частоты основного тона зависит от размеров и степени натяжения связок. [20] Кроме самой частоты основного тона оценивается ее среднеквадратическое отклонение. В таблице 1.1 представлена связь между эмоцией и изменением этих характеристик относительно безэмоциональной речи.

Таблица 1.1 – Связь характеристик частоты основного тона и эмоции

Базовая эмоция	*Изменение значений относительно нейтрального состояния*	
	ЧОТ	*СКО ЧОТ*
Радость	выше	выше
Печаль	ниже	ниже
Гнев	ниже	выше
Удивление	ниже	выше
Страх	выше	выше

Методы определения частоты основного тона можно разделить на три категории: основанные на временной динамике сигнала (*англ. time-domain*), основанные на частотной структуре (*англ. frequency-domain*) и комбинированные методы. [21]

Перед применением методов, основанных на временной динамике, сигнал предварительно фильтруют, оставляя только низкие частоты. Задаются минимальная и максимальная частоты (например, от 75 до 500 Гц). Частота основного тона не определяется для участков, содержащих негармоничную речь (паузы, шумовые звуки), поскольку это влечет за собой ошибки, которые могут распространяться на соседние фреймы при применении интерполяции или сглаживания. Длину фрейма выбирают так, чтобы в ней содержалось как минимум три периода.

В методах, основанных на частотной структуре, анализируется гармоническая структура сигнала. Одним из таких методов является кепстральный анализ. Кепстр – результат преобразования, получаемого применением обратного преобразования Фурье к логарифму спектра мощности сигнала. Иными словами, кепстр является спектром от спектра сигнала.

Алгоритм получения кепстра можно разделить на три этапа.

1. К сигналу применяется дискретное преобразование Фурье.
2. От полученного спектра мощности сигнала берется логарифм, чтобы получить лог-спектр.
3. К лог-спектру применяется обратное преобразование Фурье. На выходе получается спектр от спектра сигнала.

Гибридные методы определения имеет смысл рассмотреть на примере алгоритма YAAPT (*англ. Yet Another Algorithm of Pitch Tracking*), который считается гибридным, поскольку использует как частотную, так и временную информацию. YAAPT, как и другие алгоритмы определения ЧОТ состоит из трех этапов: препроцессирование, поиск кандидатов (возможных значений ЧОТ) и выбор наиболее вероятной траектории ЧОТ.

На этапе препроцессирования значения изначального сигнала возводят в квадрат для усиления и восстановления пиков автокорреляции. Затем по спектру преобразованного сигнала рассчитывается базовая траектория ЧОТ. Кандидаты определяются с помощью функции SHC – *от англ. Spectral Harmonics Correlation* согласно 1.1:

$$\mathrm{SHC}(t,\, f) = \sum_{f'=-\mathrm{WL}/2}^{\mathrm{WL}/2} \prod_{r=1}^{\mathrm{NH}+1} S(t,\, rf + f'), \tag{1.1}$$

где $S(t,\, f)$ — магнитудный спектр для фрейма t и частоты f, WL — длина окна (Гц), NH – число гармоник (рекомендуется [22] использовать первые три гармоники).

Далее, как для изначального сигнала, так и для преобразованного производится определение кандидатов на F0, и вместо автокорреляционной функции здесь используется функция NCCF – *от англ. Normalized Cross Correlation* (1.2):

$$\mathrm{NCCF}(m) = \frac{\sum_{n=0}^{N-m-1} x(n) \cdot x(n+m)}{\sqrt{\sum_{n=0}^{N-m-1} x^2(n) \cdot \sum_{n=0}^{N-m-1} x^2(n+m)}},\ 0 < m < M_0 \tag{1.2}$$

Далее проводится оценка всех возможных кандидатов и вычисление их веса (*англ. merit*). Вес кандидатов, полученных по аудиосигналу, от амплитуды пика NCCF и от их близости к траектории ЧОТ, определенной по спектру.

Затем для всех пар оставшихся кандидатов рассчитывается матрица цены перехода (*англ. Transition Cost*), по которой находят оптимальную траекторию ЧОТ.

1.3.1.2. Инетнсивность, темп речи и паузация

Громкость (интенсивность) речи измеряется в децибелах (дБ). Связь градаций интенсивности и эмоции представлена в таблице 1.2.

Таблица 1.2 – Связь градаций интенсивности и эмоции

Интенсивность	*Значение* (дБ)
Шепот	< 20
Значительное снижение	20. . . 40
Умеренное снижение	40. . . 50
Нормальная	50. . . 80
Умеренное повышение	80. . . 90
Значительное повышение	90. . . 110
Крик	> 110

Следующей просодической характеристикой является паузация. Короткими паузами принято считать паузы до 3 секунд, средними – от 3 до 7 секунд, длинными – свыше 7 секунд. В таблице 1.3 представлена связь характеристик паузации и эмоции.

Таблица 1.3 – Связь характеристик паузации и эмоции

Базовая эмоция	*Изменение значений относительно нейтрального состояния*	
	количество пауз	длина пауз
Радость	меньше	короче
Печаль	больше	короче
Гнев	меньше	короче
Удивление	больше	длиннее
Страх	больше	длиннее

Темпом речи называют скорость произнесения элементов речи (звуков, слогов, слов). Темп речи изменяется двумя параметрами: [23]

– числом произносимых в единицу времени элементов речи;

– средней длительностью элемента.

Повышение темпа речи осуществляется за счет сокращения длительности гласных и согласных звуков, снижение темпа достигается путем увеличения длительности гласных. Согласно исследованиям в [24], изменение темпа речи связано с проявлением говорящим эмоции. Например, проявление тревожности характеризуется ускорением темпа речи, а холодность и задумчивость – замедлением.

1.3.2. Спектральные характеристики

1.3.2.1. Мел-кепстральные коэффициенты

Основной принцип работы с человеческой речью заключается в том, что звуки, генерируемые человеком, фильтруются формой голосового тракта (язык, зубы и т.д.). Набор этих характеристик и можно представить с помощью мел-кепстральных коэффициентов (*англ. MFCC*). [25]

Шкала Мел (рисунок 1.1) соотносит воспринимаемую частоту или высоту чистого тона (мел) с фактической измеренной частотой (Гц). Люди гораздо лучше различают небольшие изменения высоты звука на низких частотах, чем на высоких. [26]

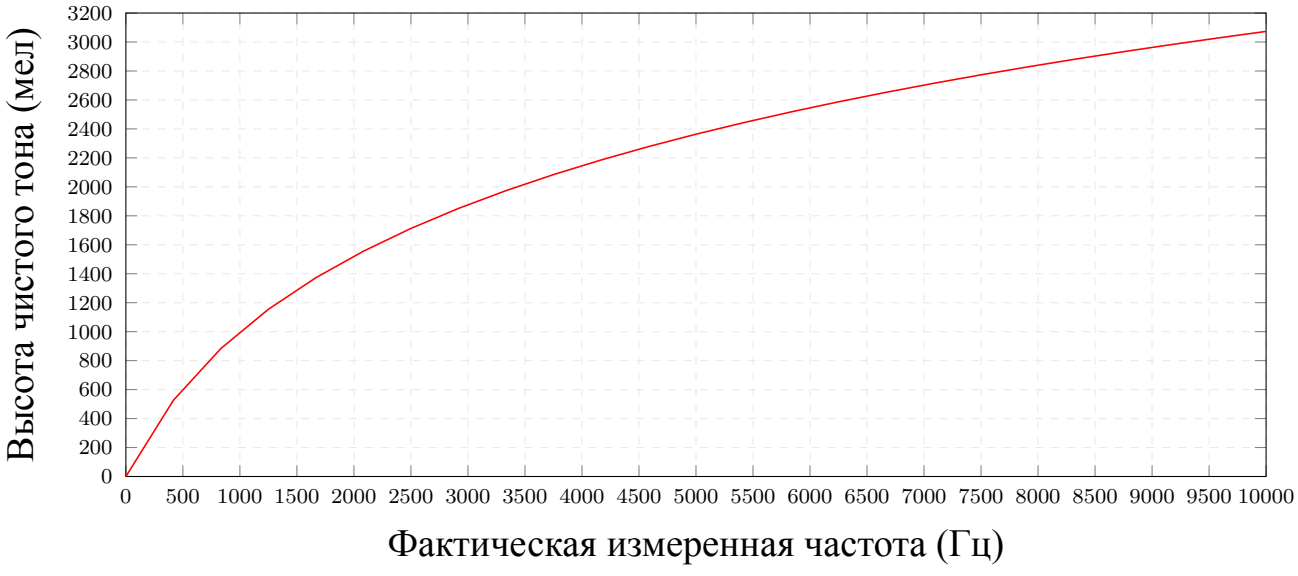

Рис. 1.1 – График зависимости частоты от мел

Вычисление мел-кепстральных коэффициентов заключается в следующем. Для каждого фрейма $x_j(n)$ выполняется дискретное преобразование Фу-

рье (1.3):

$$X_j(k) = \sum_{n=0}^{N-1} x_j(n) w(n) \exp -\frac{2\pi i}{N} kn,\ 0 \le k < N, \tag{1.3}$$

где j – номер фрейма, $w(n)$ – оконная функция Хэмминга, используемая для уменьшения утечки ДПФ на интервале конечной длительности.

Следующим шагом вычисляется банк мел-фильтров из M треугольных фильтров. Для этого треугольные фильтры умножаются на периодограмму и суммируются. Каждый треугольный фильтр моделируются с помощью функции 1.4:

$$H_m(k) = \begin{cases} 0, & k < f(m-1), \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k < f(m), \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k < f(m+1), \\ 0, & k > f(m+1). \end{cases} \tag{1.4}$$

Далее производится расчет логарифмического значения энергии компонент спектра на выходе каждого фильтра (1.5):

$$T_j(m) = \ln \sum_{k=0}^{N-1} P_j(k) H_m(k),\ 0 \le m < M. \tag{1.5}$$

Поскольку ДПФ характеристик синтезированных фильтров 1.4 взаимно пересекаются, а энергии на выходе фильтров существенно коррелируют, для вычисления MFCC необходимо использовать дискретное косинусное преобразование (1.6), чтобы устранить возникающие корреляци:

$$c_j(m) = \sum_{m=0}^{M-1} T_j(m) \cos \left(\frac{\pi n \left(m + \dfrac{1}{2} \right)}{M} \right),\ 0 \le n < M. \tag{1.6}$$

После получения $c_j(m)$, коэффициент $c_j(0)$ отбрасывается, так как он не несет информации о речи диктора и задает постоянное смещение. [17]

1.3.2.2. Энергетические и пертурбационные параметры

Выделяют следующие энергетические параметры речи: нижний уровень громкости речи (I_{lower}), верхний уровень громкости речи (I_{higher}), динамический диапазон (ΔI) и амплитуда основного тона (I_{pinch}). Энергетические параметры речи измеряются в децибелах (дБ).

Пертурбационными параметрами называют джиттер и шиммер. Существует несколько параметров их оценки. Локальный джиттер определяется согласно 1.7:

$$\text{jitter}_{\text{loc}} = \frac{1}{N-1}\sum_{i=1}^{N-1}|F_0(i) - F_0(i+1)| \Big/ \frac{1}{N}\sum_{N}^{i=1} F_0(i), \tag{1.7}$$

где $F_0(i)$ – длительность i-го периода основного тона, N – число периодов основного тона.

Другой способ определения джиттера – использование отклонения текущей длительности периода не от предыдущей, а от локально усредненного значения, которое рассчитывается на окне в 3 или 5 выборок [27]:

$$\text{jitter}_{\text{P}} = \frac{1}{N-P+1}\sum_{i=1+\frac{P-1}{2}}^{N-\frac{P-1}{2}} \left| F_0(i) - \frac{1}{P}\sum_{n=i-\frac{P-1}{2}}^{i+\frac{P-1}{2}} F_0(n) \right| \Big/ \frac{1}{N}\sum_{i=1}^{N}|F_0(i)|, \tag{1.8}$$

где P – параметр, определяющий количество периодов используемых для вычисления среднего.

Аналогично джиттеру, для вычисления шиммера существует несколько способов. Простейший – определение средней абсолютной разницы между амплитудами последовательных периодов основного тона $A(i)$, деленной на среднюю амплитуду (выражение 1.9):

$$\text{shimmer}_{\text{loc}} = \frac{1}{N-1}\sum_{i=1}^{N-1}|A(i) - A(i+1)| \Big/ \frac{1}{N}\sum_{N}^{i=1} A(i). \tag{1.9}$$

Существуют варианты, определения шиммера как относительное отклонение амплитуды от локально-усредненного значения на интервале в P выбо-

рок(1.10):

$$\mathrm{shimmer_P} = \frac{1}{N-P+1} \sum_{i=1+\frac{P-1}{2}}^{N-\frac{P-1}{2}} \left| A(i) - \frac{1}{P} \sum_{n=i-\frac{P-1}{2}}^{i+\frac{P-1}{2}} A(n) \right| \Big/ \frac{1}{N} \sum_{i=1}^{N} |A(i)| \,. \tag{1.10}$$

Такая оценка более точна чем 1.9, поскольку на шиммер влияет постепенный равномерный (естественный) спад интенсивности голоса, создающий эффект «дрейфа» амплитуды сигнала. [28]

1.3.2.3. Частоты первых формант речевого сигнала

Форманты возникают под влиянием резонаторов речевого аппарата, поэтому их высотное положение не зависит от основного тона, но зависит от произносимого звука. При произнесении некоторых звуков речи (в основном гласных) происходит усиление обретонов на определенной частоте: например, при произнесении гласной ”у”характерно усиление частичных тонов от 200 до 400 Герц, а для гласной ”о от 400 до 600 Герц.

Определение формант возможно с помощью кепстрального анализа. Речь можно представить как результат свертки инициирующего сигнала гортанных импульсов с частотной характеристикой голосового тракта, выступающего в роли линейного фильтра (1.11):

$$x(t) = s(t) * h(t), \tag{1.11}$$

где $s(t)$ - инициирующий сигнал, $h(t)$ – частотная характеристика голосового тракта. Представляя сигнал во временной области согласно 1.12:

$$X(\omega) = S(\omega) \cdot H(\omega), \tag{1.12}$$

при переходе к лог-спектру (1.13) и кепстру (1.14) можно получить следующие выражения:

$$\log X(\omega) = \log S(\omega) + \log H(\omega) \tag{1.13}$$

$$X(\bar{\omega}) = S(\bar{\omega}) + H(\bar{\omega}) \tag{1.14}$$

То есть свертка сигналов во временной области эквивалентна их умножению в частотной области или сложению в лог-спектре и кепстре. Это свойство кепстра позволяет разделить инициирующий сигнал и частотную характеристику голосового тракта. При этом информация о тембре и формантах будет находиться на низких quefrency-значениях (анаграмма от англ. *frequency*, дословно – сачтотах) кепстра, а информация о инициирующем сигнале - на высоких.

Чтобы извлечь тот или иной компонент сигнала, нужно применить т.н. liftering (анаграмма от англ. *filtering*, фильтрация) - аналог частотных фильтров для quefrency-области. Для получения информации о формантах можно использовать следующий фильтр низких quefrency-значений:

$$l(n) = \begin{cases} 1, n < \tau \\ -1, n \geq \tau \end{cases} \tag{1.15}$$

Для мужской речи значение τ выбирают в диапазоне от 4 до 8 мс, так как среднее значение частоты мужской речи находится в диапазоне от 128 до 270 Гц. Для женской речи, из-за более высокого среднего значения частоты (256-310 Гц), диапазон, на котором расположены форманты в кепстре, может пересекаться с диапазоном инициирующего сигнала, что мешает однозначному разделению компонент речи.

1.4. Классификаторы, используемые в анализе речи

Автоматическое определение эмоций происходит с помощью классификатора – обученной модели, которая после обучения сможет определять эмоции по записям человеческой речи. Обучение классификатора происходит с использованием одного набора данных на одном языке. Существует множество классификаторов, используемых в схожих с исследуемой задачах – модель гауссовых смесей, метод опорных векторов, метод k-ближайших соседей. Однако, наиболее популярны при распознавании эмоций модели – это скрытая марковская модель (*англ. Hidden Markov Model, HMM*) и искусственная нейронная сеть (*англ. Artificial Neural Networks, ANN*).

1.4.1. Скрытая марковская модель

Скрытые марковские модели предоставляют более подходящий и более мощный инструмент моделирования в ситуациях когда состояния не являются непосредственно наблюдаемыми. Более детальное объяснение скрытых марковских моделей имеет смысл начать с определения марковских цепей.

Марковская цепь – это конечный автомат, имеющий дискретное число состояний $q_1, \ldots, q_n$ и существует вероятность перехода из состояния q_i в другое состояние q_j: $P\left(S_t = q_j | S_{t-1} = q_i\right)$. Цепь может находиться в состоянии q_i в любой момент времени t, поскольку вермя дискретно. Также согласно марковскому свойству, вероятность следующего состояния зависит только от вероятности предыдущего.

Скрытая марковская модель – это марковская цепь, в которой состояния не являются непосредственно наблюдаемыми. Такую модель можно объяснить как двойной стохастический процесс: скрытый стохастический процесс, который невозможно наблюдать напрямую и процесс, который создает последовательность наблюдений с учетом первого процесса.

При применении скрытой марковской модели при классификации выделяют три основных задачи.

1. *Задача вычисления оценки*. В рассматриваемой модели необходимо определить оценку вероятности последовательности наблюдений.

2. *Задача определения оптимальной последовательности*. С учетом модели и конкретной последовательности наблюдений необходимо определить оценку наиболее вероятной последовательности состояний, которая создает эти наблюдения.

3. *Задача обучения параметров*. С учетом количества последовательностей наблюдений необходимо отрегулировать параметры модели.

Для того, чтобы формализовать представленные задачи, следует ввести следующие обозначения. Модель $\lambda = (A,\ B,\ \pi)$ состоит из матрицы перехода A, матрицы наблюдаемых значений B и начального распределения π. Последовательность наблюдаемых значений D выбирается из алфавита (множества значений скрытых параметров) V.

Суть первой задачи заключается в определении вероятности последовательности наблюдений D по параметрам данной модели $\lambda = (A, B, \pi)$. Для вычисления вероятности последовательности наблюдений можно вычислить ее оценку для конкретной последовательности состояний, а затем прибавить вероятности для всех возможных последовательностей состояний согласно 1.16:

$$P(D|\lambda) = \sum_{Q} P(D|\, Q,\, \lambda) p(Q|\, \lambda) \tag{1.16}$$

Вторая задача заключается в том, чтобы по модели λ и последовательности D найти оптимальную последовательность состояний Q. Третья задача – главная, заключается в оптимизации параметров модели $\lambda = (A, B, \pi)$ таким образом, чтобы минимизировать $p(D|\lambda)$ при данных D, т.е. найти модель максимального правдоподобия.

1.4.2. Искусственная нейронная сеть

Нейронная сеть – математическая модель, построенная на принципах функционирования биологических нейросетей (сетей нервных клеток живого организма). [29] Основная парадигма нейронных сетей – это формирование решения из множество простых элементов, подобных нейронам. Эти элементы образуют граф с взвешенными синаптическими связями. Искусственная нейронная сеть обладает следующими свойствами: [30]

– параллельность – в любой момент времени в активном состоянии могут находиться несколько процессов;

– распределенность – каждый из процессов может независимо обрабатывать локальные данные;

– свойство самообучения и подстройки своих параметров при изменении профиля данных.

Единицу, выполняющую вычисления в нейронной сети, называют нейроном. Нейроны обрабатывают входной сигнал и отправляют его дальше по сети. Нейрон представляет собой некую функцию от линейной комбинации всех своих входных сигналов. Основная функция нейрона - сформировать выходной сигнал y в зависимости от сигналов $x_1, \ldots, x_N$, поступающих на его входы.

Входные сигналы обрабатываются адаптивным сумматором (1.17):

$$\sum_{i=1}^{N} w_i x_i - T, \tag{1.17}$$

где T – порог нейрона, $w_1, \ldots, w_N$ – знаки весов синапсов. Выходной сигнал поступает в нелинейный преобразователь F с некоторой функцией активации, после чего результат подается на выход (в точку ветвления).

Синапс – связь между нейронами, причём каждый синапс имеет свой вес. Благодаря этому входные данные видоизменяются при передаче. Во время обработки переданная синапсом информация с большим показателем веса станет преобладающей.

Схема классификации на основе нейронных сетей включает в себя следующие шаги.

1. На входной слой нейронов происходит поступление определённых данных.

2. Информация передаётся с помощью синапсов следующему слою, причём каждый синапс имеет собственный коэффициент веса, а любой следующий нейрон способен иметь несколько входящих синапсов. Данные, полученные следующим нейроном – это сумма всех данных для нейронных сетей, которые перемножены на коэффициенты весов.

3. Полученное в итоге значение подставляется в функцию активации (для нормализации входных данных), в результате чего происходит формирование выходной информации.

4. Информация передаётся дальше до тех пор, пока не дойдёт до конечного выхода.

1.5. Постановка задачи

В настоящее время ряд задач, связанных с распознаванием эмоций в речи решается с использованием нейронных сетей или статистических классификаторов. В настоящей работе проектируется метод решения этой задачи с использованием статистического классификатора, а именно – скрытой марков-

ской модели. На рисунке 1.2 представлена IDEF0-диаграмма нулевого уровня решаемой задачи.

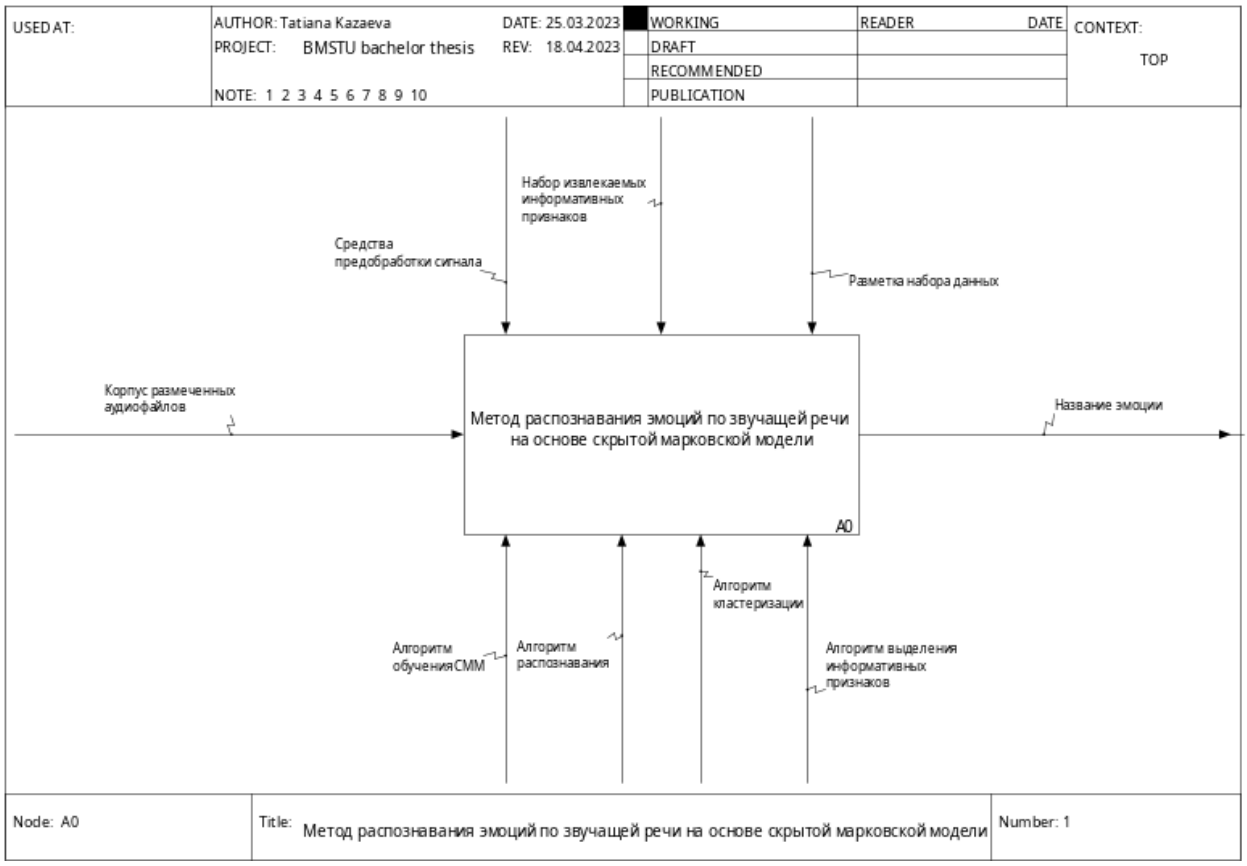

Рис. 1.2 – IDEF0-диаграмма нулевого уровня

Набор данных должен быть подготовлен к использованию классификатором – а именно, должно быть установлено однозначное соответствие аудиофайл-эмоция согласно представленной в наборе разметке. Под исследуемую задачу по разметке и по языку подходят следующие наборы: RUSLANA и DUSHA.

Из аудиозаписей должны быть выделены информативные признаки. Для обучения классификатора эти признаки должны быть поделены на кластеры. Обученный классификатор используется для распознавания эмоций.

2 Конструкторский раздел

2.1. Общая схема метода

Задача распознавания эмоций по звучащей речи сводится к соотношению исходных данные на входе (аудиозаписи звучащей речи) к определенному классу на выходе (виду эмоции). На рисунке 2.1 представлена IDEF0-диаграмма первого уровня решаемой задачи.

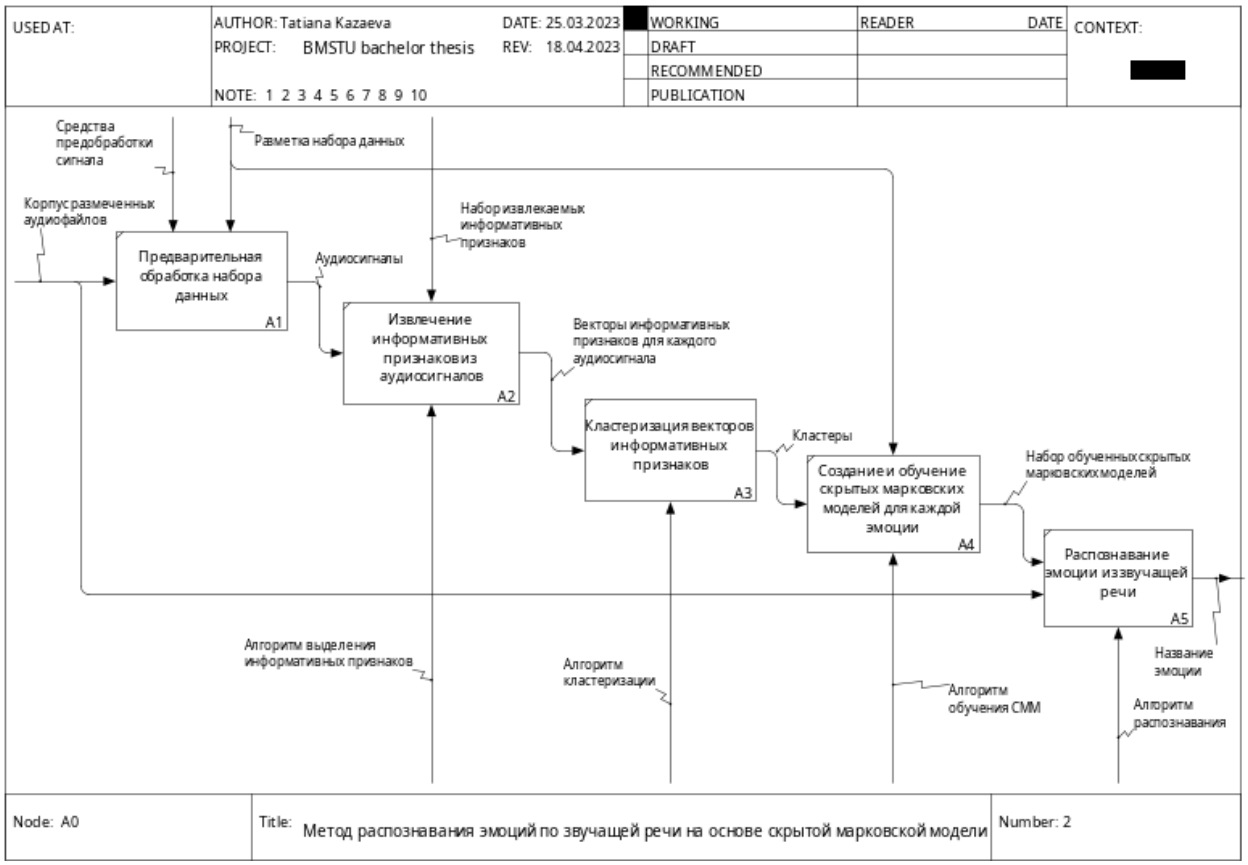

Рис. 2.1 – IDEF0-диаграмма первого уровня

Прежде всего набор данных должен быть предварительно обработан (блок A1). Предварительная обработка включает в себя разделение набора данных на тренировочную (70-80% набора) и тестовую (20-30% набора) выборки. Следующий этап решения задачи (блок A2) – это извлечение информативных признаков из речевого сигнала. Поскольку сигнал должен быть разбит на кадры (фреймы) небольшой длительности, использование просодических признаков не имеет смысла, следовательно, будут использованы спектральные признаки речевого сигнала. Далее необходимо сократить размерность вектора информативных признаков. Для этого производится кластеризация векторов признаков

сигналов (блок А3). Этап создания и обучения скрытой марковской модели для каждой эмоции (блок А4) включает в себя определение количества состояний модели, а также вероятностей перехода между состояниями и вероятностей наблюдения признаков в каждом состоянии. В качестве состояний модели будут использованы порядковые номера выделенных кластеров. Обученную марковскую модель можно использовать для распознавания эмоций в звучащей речи (блок А5).

2.2. Проектирование ключевых модулей системы

2.2.1. Формирование вектора информативных признаков

Самым надежным решением при распознавании эмоций в речи принято считать использование мел-кепстральных коэффициентов в качестве информативных признаков [31]. В данной работе решено использовать первые 13 мел-кепстральных коэффициентов.

Формирование вектора информативных признаков, состоящего из 13-ти первых мел-кепстральных коэффициентов представлено на рисунке 2.2.

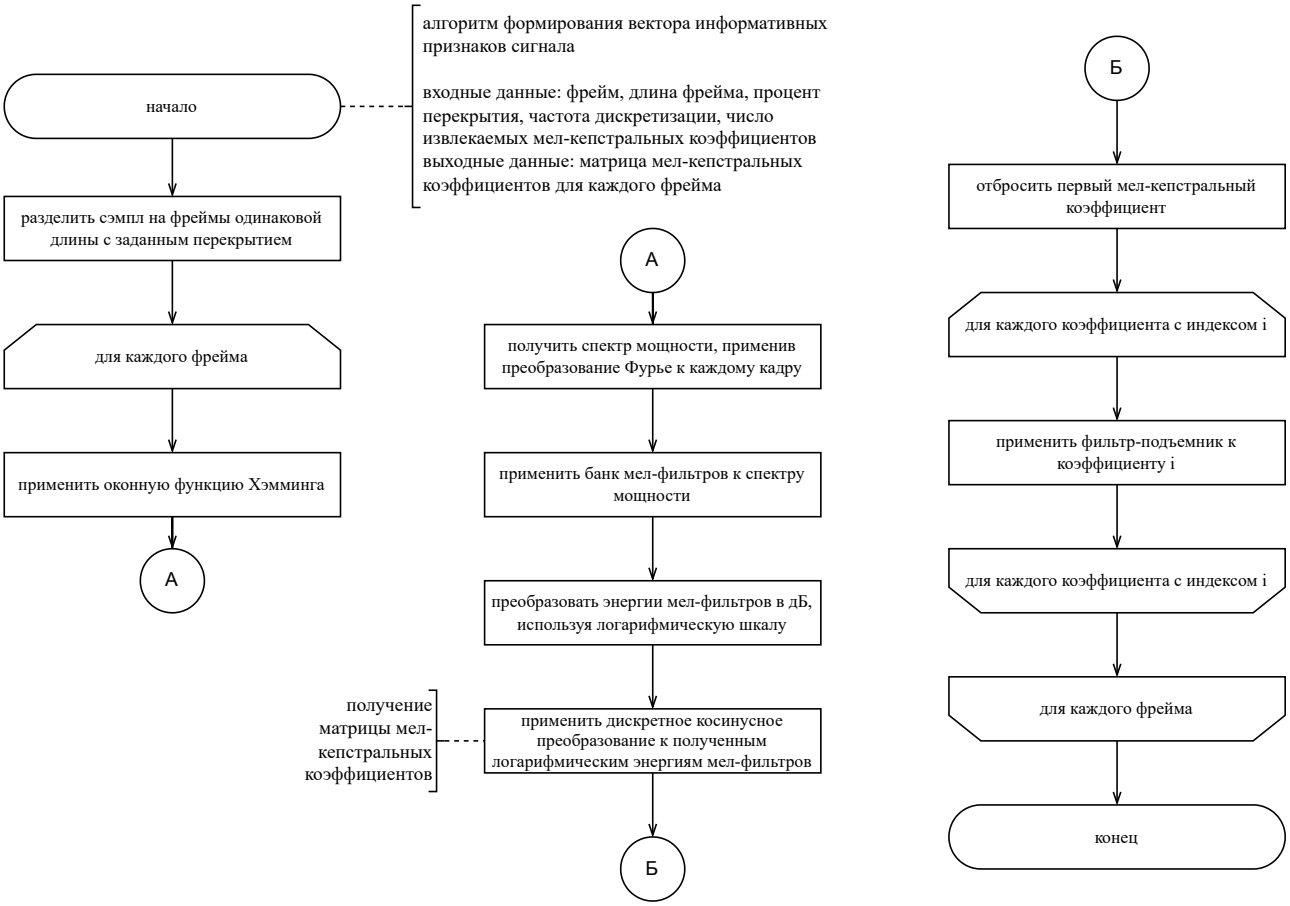

Рис. 2.2 – Алгоритм формирования вектора информативных признаков

Фильтр-подъемник (lifter) применяется для усиления высокочастотных компонентов сигнала, которые могут быть утрачены в процессе преобразования сигнала в мел-кепстральные коэффициенты. Применение фильтра-подъемника происходит согласно 2.1

$$1 + \frac{\text{lifter}}{2} \cdot \frac{\sin\left(\frac{\pi i}{\text{lifter}}\right)}{1}, \tag{2.1}$$

где i - индекс коэффициента MFCC, а lifter - значение параметра фильтра-подъемника.

2.2.2. Кластеризация

Алгоритм k-means – это неиерархичный метод неконтролируемого обучения. Он позволяет разделить произвольный набор данных на заданное число кластеров так, что объекты внутри одного кластера были достаточно близки друг к другу, а объекты разных не пересекались. Цель этого алгоритма — объединить в группы сходные данные по некоторым заданным критериям. Чаще всего при кластеризации используются меры расстояния. В данной работе в качестве меры расстояния будет реализовано Евклидово (квадратичное) расстояние согласно 2.2:

$$\text{dist}(p,\ q) = \sqrt{(p-q)^2}, \tag{2.2}$$

где p, q – точки вектора входных данных.

В качестве входных данных алгоритм принимает массив 13-ти мерных узлов, которые представляют собой значения мел-кепстральных коэффициентов, наблюдаемых на каждом фрейме каждого сэмпла. В качестве выходных данных – массив 13-ти мерных узлов, являющихся центроидами каждого кластера, количество итераций для корректировки и количество кластеров, которое определяется эвристически. Схема алгоритма кластеризации представлена на рисунке 2.3.

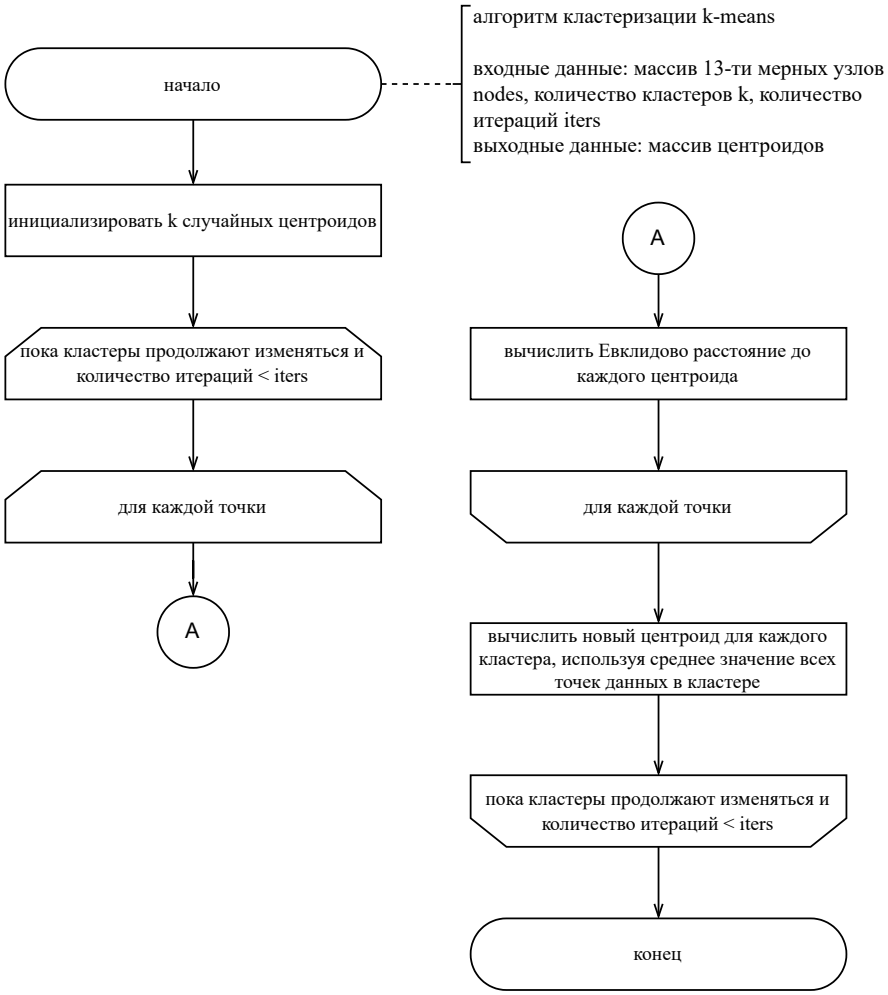

Рис. 2.3 – Алгоритм кластеризации k-means

2.2.3. Создание и обучение скрытых марковских моделей

Обучение скртыой марковской модели – это определение параметров $\lambda = \{A, B, \pi\}$ с учетом количества последовательностей наблюдений $\{O = O_1, \ldots, O_n\}$. Для обучения скрытой марковской модели используется алгоритм Баума — Велша. Стоит отметить, что он применим только в том случае, если предварительно определено количество состояний и наблюдений. Пусть γ – условная вероятность нахождения в определенном состоянии q_i в с учетом последовательности наблюдений (2.3):

$$\gamma_i = P(s_t = q_i \,|\, O,\, \lambda) = \frac{P(s_t = q_i,\, O \,|\, \lambda)}{P(O)}. \tag{2.3}$$

Далее следует ввести переменную α, обозначающую вероятность частичной последовательности наблюдений до момента времени t, находящаяся в состоянии

q_i в момент времени t, согласно 2.4:

$$\alpha_i = P(O_1, O_2, \ldots, O_t,\ S_t = q_i \mid \lambda), \tag{2.4}$$

и переменную β, вероятность частичной последовательности наблюдений от $t+1$ до T, при нахождении в состоянии q_i в момент времени t (2.5):

$$\beta_i = P(O_{t+1}, O_{t+2}, \ldots, O_T,\ S_t = q_i \mid \lambda), \tag{2.5}$$

С учетом этих обозначений следует ввести переменную ξ, обозначающую вероятность перехода из состояния i в момент времени t в состояние j в момент времени $t+1$ с учетом последовательности наблюдений O согласно 2.6:

$$\xi_t(i,\ j) = \frac{\alpha_t(i) a_{i,\,j} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O)}. \tag{2.6}$$

С учетом переменных α и β (2.4 и 2.5) задать ξ удобнее согласно 2.7:

$$\xi_t(i,\ j) = \alpha_t(i) a_{i,\,j} b_j(O_{t+1} \beta_{t+1}(j) \Big/ \sum_i \sum_j \alpha_t(i) a_{i,\,j} b_j(O_{t+1} \beta_{t+1}(j) \tag{2.7}$$

С учетом введенных обозначений, Баума—Велша можно разделить на 5 этапов.

1. Выполнение прямого прохода, в результате которого вычисляются вероятности наблюдаемых последовательностей до каждого момента времени и вероятности переходов из одного скрытого состояния в другое.

2. Выполнение обратного хода, в результате которого вычисляются апостериорные вероятности скрытых состояний в каждый момент времени.

3. оценка априорных вероятностей – количество случаев нахождения в состоянии i в момент времени t согласно 2.8:

$$\pi_i = \gamma_1(i). \tag{2.8}$$

4. Оценка вероятностей переходов – количество переходов из состояния i в состояние j по отношению к количеству случаев нахождения в состоянии

i согласно 2.9:

$$a_{i,j} = \frac{\sum_{t=1}^{T} \gamma_j(t) 1(v(t) = k)}{\sum_{t=1}^{T} \gamma_j(t)}. \tag{2.9}$$

5. Оценка вероятностей наблюдений – количество случаев нахождения в состоянии j при количестве наблюдений k по отношению к количеству случаев нахождения в состоянии j согласно 2.10:

$$b_{j,k} = \frac{\sum_{t=1}^{T} \gamma_j(t) 1(v(t) = k)}{\sum_{t=1}^{T} \gamma_j(t)} \tag{2.10}$$

Выполнение прямого прохода осуществляется используя алгоритм прямого хода (формализован согласно 2.1), который состоит из трех основных частей: инициализация, индукция и завершение. На этапе инициализации определяются переменные α для всех состояний в начальный момент времени. На этапе индукции вычисляются значения $\alpha_{t+1}(i)$ по значениям $\alpha_t(i)$. На этапе завершения вычисляется значение $P(O \mid \lambda)$ путем суммирования всех значений α_T.

Алгоритм 2.1: Алгоритм прямого хода

Исходные параметры: Скрытая марковская модель λ; Последовательность наблюдений O; количество состояний N; Количество наблюдений T

$i \leftarrow 0$; **цикл** $i < T$ **выполнять**
  $\alpha_1(i) = \pi_i b_i(O_1)$; // инициализация
конец цикла
$t \leftarrow 1$; **цикл** $t < T$ **выполнять**
  $j \leftarrow 0$; **цикл** $j < N$ **выполнять**
    $\alpha_t(j) = \sum_i \alpha_{t-1}(i) a_{i,j} b_j(O_t)$ // индукция
  конец цикла
конец цикла
$P(O) = \sum_i \alpha_T(i)$

Алгоритм, с помощью которого осуществляется обратный ход, работает анало-

гично, но в нем инициализация и индукция начинаются с конца последовательности.

2.2.4. Определение эмоции из аудиосигнала

2.3. Описание используемого набора данных

2.3.1. Разметка и структура набора

Для обучения классификатора было решено использовать набор данных DUSHA [14], содержащий записи эмоциональной речи. Набор разделен на два домена – аудиоматериалы, собранные с помощью краудсорсинга (*англ. «Croud»*) и выдержки из русскоязычных подкастов (*англ. «Podcast»*).

При разметке эмоциональных наборов данных существует сложность в неоднозначности интерпретации эмоции аннотаторами. В наборе данных DUSHA эта проблема решается привлечением к оцениванию двух независимых экспертных групп. В наборе присутствуют только те записи, для которых оценка обоих групп была консистентна. В разметке присутствуют следующие классы эмоций:

- **нейтраль**;
- **позитив**: текст требовалось произносить с улыбкой или смехом, стараться делать выраженные ударения на позитивно окрашенных словах;
- **грусть**: произносить текст требовалось приглушенным голосом, меланхолично;
- **злость или раздражение**: текст требовалось произнести с криком или сквозь зубы, и, аналогично позитиву, стараться делать выраженные ударения на негативно окрашенных словах.

В домене «Crowd» тексты были искусственно сгенерированы на основе записей общения с голосовым ассистентом. Затем они были озвучены двумя способами: эмоционально, с той эмоцией, которую показал на этой записи классификатор BERT [32] и безэмоционально. Домен «Podcast» содержит уже не имитацию эмоции, а естественную речь. Все аудиозаписи были сделаны на профессиональные микрофоны, качество аудио было унифицировано до 16кГц.

Аннототоры производили разметку, опираясь исключительно на звуковую дорожку, без учета произнесённого в ней текста.

2.3.2. Содержание набора данных

Поскольку разметку осуществляли несколько аннотаторов, требовалась агрегарция разметки. Она производилась по методу Дэвида — Скина с порогом 0.9, выбранным эмпирически [14]. Объём данных после агрегации с разбивкой по подмножествам приведён в таблице 2.1.

Таблица 2.1 – Объём данных после агрегации

Домен	Тренировочная выборка		Тестовая выборка	
	Файлы (шт.)	Время	Файлы (шт.)	Время
Crowd	147057	184 ч. 21 мин.	13867	18 ч. 17 мин.
Podcast	78810	70 ч. 08 мин.	10591	09 ч. 24 мин.
Всего	225867	254 ч. 29 мин.	24458	27 ч. 41 мин.

В агрегированном наборе данных имеются следующие поля разметки:

- **audio_path:** путь к аудиофайлу;
- **emotion:** эмоция, которую указал разметчик;
- **speaker_text:** текст, который произнёс диктор (присутствует только в домене Crowd);
- **speaker_emo**: эмоция, которую выражал диктор (присутствует только в домене Crowd);
- **source_id**: уникальный идентификатор диктора или подкаста.

2.4. Проектирование отношений сущностей

На каждом этапе решения поставленной задачи формируется большой объем связанных структурированных данных, который необходимо хранить и дополнять. Для хранения этого набора решено использовать реляционную базу данных. На рисунке 2.4 представлена диаграмма сущностей базы данных в нотации Чена.

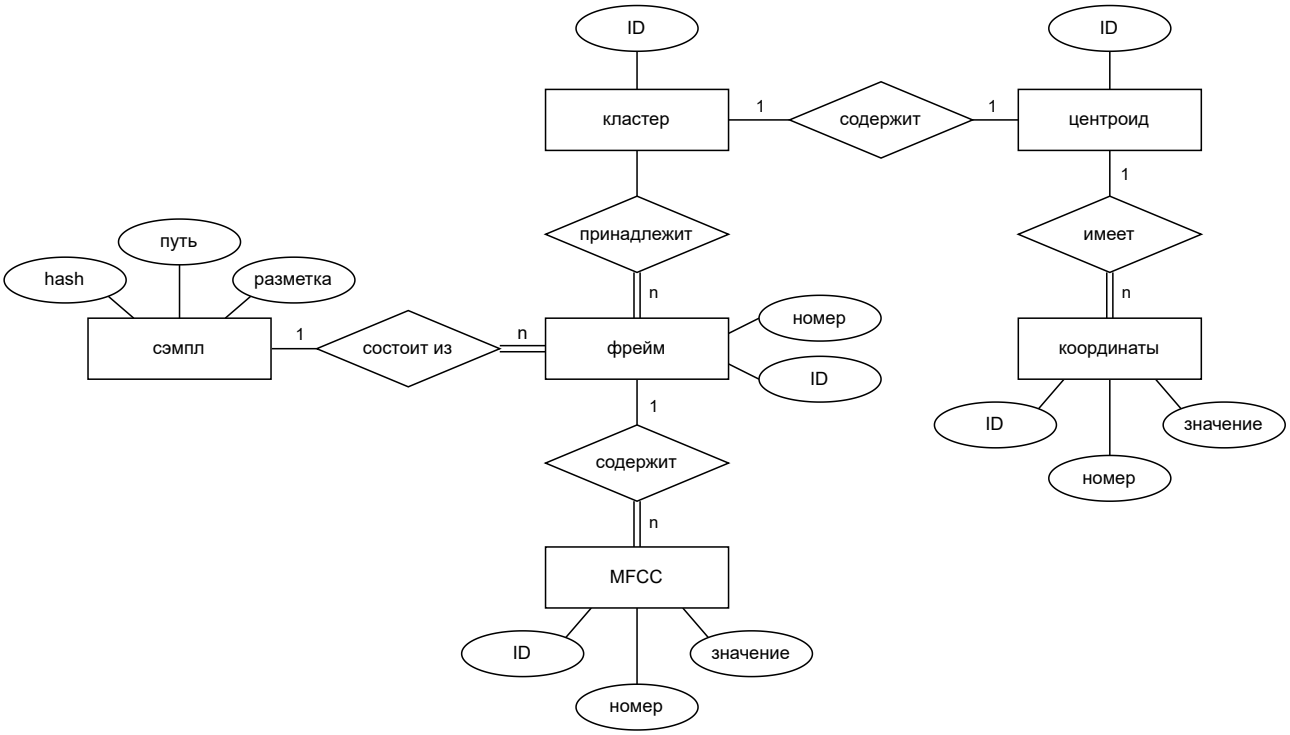

Рис. 2.4 – Диаграмма сущностей базы данных в нотации Чена

Сущность «сэмпл» представляет собой информацию о WAV-файле, который хранится в файловой системе. Сущность содержит поля, необходимые для ее обработки: уникальный хэш для идентификации сущности, абсолютный путь в файловой системе компьютера и разметка, указанная в корпусе. Звуковые дорожки разделены на небольшие фрагменты для дальнейшего анализа. Эти фрагменты представлены сущностью «фрейм». Для работы с фреймами необходимо хранить информацию о его порядковом номере в сэмпле и идентификатор.

Полученный набор фреймов используется в качестве входных данных для кластеризации. В результате кластеризации каждому фрейму присвоен кластер. Сущность базы данных, хранящая кластер, содержит уникальный идентификатор и информацию о своем центроиде. Центроид, в свою очередь, также содержит идентификатор и набор координат.

2.5. Проектирование клиентского приложения

2.6. Вывод

3 Технологический раздел

3.1. Выбор средств программной реализации

3.1.1. Выбор языка программирования

3.1.2. Выбор СУБД

3.1.3. Модули программного обеспечения

3.1.4. Модуль получения вектора информативных признаков сигнала

3.1.5. Модуль формирования вектора информативных признаков

3.1.6. Модуль кластеризации

3.1.7. Модуль создания и обучения скрытых марковских моделей

4 Исследовательский раздел

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Ekman P.* Universals and Cultural Differences in Facial Expression of Emotion // Nebraska Symposium on Motivation. Vol. 19. 1972.
2. *Ekman P.* An Argument for Basic Emotions // Cognition and Emotion. 1992.
3. *Plutchik R.*, *Kellerman H.* Emotion: Theory, Research, and Experience: Vol. 1. Theories of Emotion. // London: Academic Press. 1980.
4. Affectica. [Электронный ресурс], URL: https://www.affectiva.com/ (дата обращения: 30.9.2022).
5. *Вундт В.* Психология душевных волнений // Психология эмоций. 1984.
6. RECOLA. [Электронный ресурс], URL: https://diuf.unifr.ch/main/diva/recola/ (дата обращения: 1.10.2022).
7. Sentic Computing for social media marketing / E. Cambria, M. Grassi, A. Hussain, C. Havasi // Multimedia Tools and Applications - MTA. 2012. DOI: 10.1007/s11042-011-0815-0.
8. *Russell J.* Core Affect and the Psychological Construction of Emotion // Psychological Review. 2003.
9. RAVDESS Emotional speech audio. [Электронный ресурс], URL: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio (дата обращения: 15.3.2023).
10. Combining frame and turn-level information for robust recognition of emotions within speech / B. Vlasenko, B. Schuller, A. Wendemuth, G. Rigoll //. 01.2007. С. 2249—2252.
11. EmoDB Dataset. [Электронный ресурс], URL: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb (дата обращения: 15.3.2023).
12. Toronto emotional speech set (TESS). [Электронный ресурс], URL: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess (дата обращения: 15.3.2023).
13. *Makarova V.*, *Petrushin V. A.* RUSLANA: A database of Russian emotional utterances // Seventh international conference on spoken language processing. 2002.

14. Large Raw Emotional Dataset with Aggregation Mechanism / V. Kondratenko, A. Sokolov, N. Karpov, O. Kutuzov, N. Savushkin, F. Minkin // arXiv preprint arXiv:2212.12266. 2022.

15. Русская фонетика. Интонационные конструкции. Интонация. [Электронный ресурс], URL: http://www.philol.msu.ru/~fonetica/index1.htm (дата обращения: 15.3.2023).

16. Русская фонетика. Интонационные конструкции. Перечень ИК. [Электронный ресурс], URL: http://www.philol.msu.ru/~fonetica/intonac/ik/ik1.htm (дата обращения: 15.3.2023).

17. Метод автоматической классификации эмоционального состояния диктора по голосу / К. Кошеков, В. Кобенко, Р. Анаятова, А. Савостин, А. Кошеков // Динамика систем, механизмов и машин. 2020. Т. 8, № 4. С. 51—59.

18. Лекция 6. Признаки. Кепстральные коэффициенты. MFCC. [Электронный ресурс], URL: https://logic.pdmi.ras.ru/~sergey/oldsite/teaching/asr/notes-06-features.pdf (дата обращения: 17.2.2023).

19. *Киселев В.*, *Давыдов А.*, *Ткаченя А.* Система определения эмоционального состояния диктора по голосу. 2012.

20. *Рабинер Л.*, *Гоулд Б.* Теория и применение цифровой обработки сигналов. Рипол Классик, 1978.

21. Pitch extraction and fundamental frequency: History and current techniques / D. Gerhard [и др.]. Department of Computer Science, University of Regina Regina, SK, Canada, 2003.

22. *Zahorian S. A.*, *Dikshit P.*, *Hu H.* A spectral-temporal method for pitch tracking // Ninth International Conference on Spoken Language Processing. 2006.

23. *Ярцева В. Н.* Лингвистический энциклопедический словарь. Советская энциклопедия, 1990.

24. Распознавание эмоций по характеристикам речевого сигнала (лингвистический, клинический, информационный аспекты) / Л. П. Прокофьева, И. Л. Пластун, Н. В. Филиппова, Л. Ю. Матвеева, Н. С. Пластун // Сибирский филологический журнал. 2021. № 2. С. 325—336.

25. Первичный анализ речевых сигналов. [Электронный ресурс], URL: https://alphacephei.com/ru/lecture1.pdf (дата обращения: 17.2.2023).

26. *Chauhan P.*, *Desai N.* Mel Frequency Cepstral Coefficients (MFCC) based speaker identification in noisy environment using wiener filter // Proceedings of International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE 2014). 2014.

27. *Farrús M.*, *Hernando J.*, *Ejarque P.* Jitter and shimmer measurements for speaker recognition // 8th Annual Conference of the International Speech Communication Association; 2007 Aug. 27-31; Antwerp (Belgium).[place unknown]: ISCA; 2007. p. 778-81. International Speech Communication Association (ISCA). 2007.

28. *Baken R. J.*, *Orlikoff R. F.* Clinical measurement of speech and voice. Cengage Learning, 2000.

29. Нейронные сети. [Электронный ресурс], URL: https://logic.pdmi.ras.ru/~sergey/oldsite/teaching/asr/notes-11-neural.pdf (дата обращения: 9.3.2023).

30. *Дюк В.*, *Самойленко А.* Data Mining: учебный курс. СПб.: Питер, 2001.

31. *Lanjewar R. B.*, *Chaudhari D.* Speech emotion recognition: a review // International Journal of Innovative Technology and Exploring Engineering (IJITEE). 2013. Т. 2, № 4. С. 68—71.

32. Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // arXiv preprint arXiv:1810.04805. 2018.

ПРИЛОЖЕНИЕ А