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 18 с., 1 рис., 0 табл., 15 ист., 0 прил.

СОДЕРЖАНИЕ

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

3) **Периодограмма** – функция от частоты, которая показывает оценку спектральной плотности сигнала.

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

ВВЕДЕНИЕ

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. Представление адудиосигнала

речеобразование, представление сигнала, дискретизация

Результатом дискретизации является набор измерений $s(n)$, который был получен в момент времени $\Delta t \cdot n$ значений непрерывного сигнала.

Задача распознавания эмоций решается непосредственно по оцифрованному сигналу. Выделяется два этапа решения задачи:

– выделение и отбор информативных признаков;

– классификация (сопоставление признаков).

1.3. Выделение информативных признаков

Важной особенностью речевого сигнала является его условная стационарность на небольших промежутках (от 20 до 40 мс). [9] По этой причине для оцифровки сигнал разделяется на фреймы. Деление происходит таким образом, чтобы каждая точка перекрывалась дважды. [10]

Однако, 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: [11]

– 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.

Информативные признаки эмоциональной речи можно разбить на две категории: спектральные (линейные спектральные частоты, кепстральные коэффициенты линейной шкалы частот, кепстральные коэффициенты мел-шкалы частот) и лингвистические (мелодика речи, интенсивность, ритм, тембр, сила голоса). [12] Системы голосового детектирования эмоционального состояния, работающие со спонтанной речью, могут комбинировать акустические и лингвистические информативные признаки.

1.3.1. Мел-кепстральные коэффициенты

Для представления огибающей спектра, которой описывается форма голосового тракта были введены мел-кепстральные коэффициенты (*англ. MFCC*). [13]

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. [14]

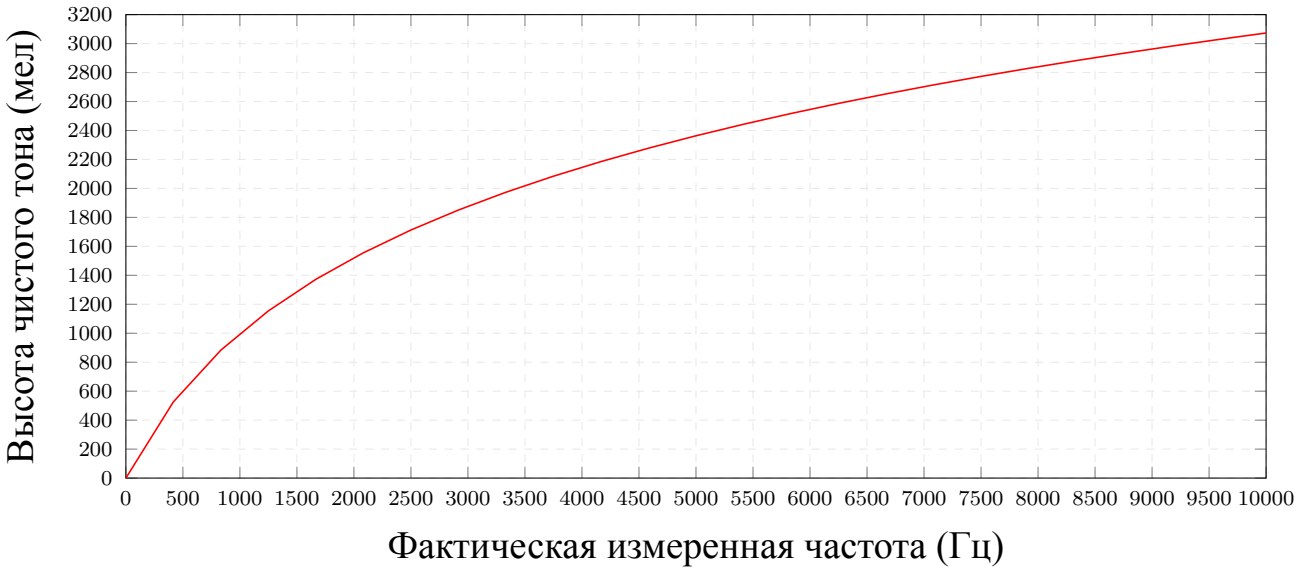

Рис. 1.1 – График зависимости частоты от мел

Вычисление мел-кепстральных коэффициентов заключается в следующем. Для каждого фрейма $x_j(n)$ выполняется дискретное преобразование Фурье (1.1):

$$X_j(k) = \sum_{n=0}^{N-1} x_j(n)w(n)\exp -\frac{2\pi i}{N}kn,\ 0 \le k < N, \tag{1.1}$$

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 функ-

ции 1.2:

$$H_m(k) = \begin{cases} 0, & k < f(m-1), \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k < f(m), \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k < f(m+1), \\ 0, & k > f(m+1). \end{cases} \tag{1.2}$$

Далее производится расчет логарифмического значения энергии компонент спектра на выходе каждого фильтра (1.3):

$$T_j(m) = \ln \sum_{k=0}^{N-1} P_j(k) H_m(k),\ 0 \le m < M. \tag{1.3}$$

Поскольку ДПФ характеристик синтезированных фильтров 1.2 взаимно пересекаются, а энергии на выходе фильтров существенно коррелируют, для вычисления MFCC необходимо использовать дискретное косинусное преобразование (1.4), чтобы устранить возникающие корреляци:

$$c_j(m) = \sum_{m=0}^{M-1} T_j(m) \cos\left(\frac{\pi n \left(m + \dfrac{1}{2} \right)}{M} \right),\ 0 \le n < M. \tag{1.4}$$

После получения $c_j(m)$, коэффициент $c_j(0)$ отбрасывается, так как он не несет информации о речи диктора и задает постоянное смещение. [9]

1.3.2. Кепстральные коэффициенты линейного предсказания

Кепстральные коэффициенты используются для оценки периода основного тона, формант и других информативных признаков эмоциональной речи.

Кепстр может быть получен путем линейного предсказания (*англ. LP*). Смысл анализа на основе линейного предсказания заключается в том, что участок речевого сигнала n можно аппроксимировать линейной комбинацией p

предыдущих участков сигнала согласно 1.5: [15]

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3) + \cdots + a_p s(n-p), \tag{1.5}$$

где $\{a_i\}_{i=1}^{p}$ – коэффициенты линейного предсказания, считаются постоянными на протяжении времени фрейма. Эти коэффициенты используются для аппроксимации участка речевого сигнала. Разница между предсказанным и действительным участком сигнала называется погрешностью предсказания и вычисляется согласно 1.6

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k), \tag{1.6}$$

откуда видно, что погрешность предсказания представляет собой сигнал на выходе системы с передаточной функцией (1.7):

$$A(z) = 1 - \sum_{k=1}^{p} a_k z. \tag{1.7}$$

Коэффициенты $\{a_i\}_{i=1}^{p}$ можно получить, минимизируя кратковременную энергию погрешности предсказания, вычисляемую согласно 1.8:

$$E_n = \sum_{m} \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right], \tag{1.8}$$

где $s_n(m)$ – сегмент речевого сигнала, выбранный в окрестности фрейма n. Минимизировать E_n следует путем вычисления $\dfrac{\delta E_n}{\delta a_i} = 0,\ i = 1, 2, \ldots, p$. После нахождения $a_k s$, кепстральные коэффициенты вычисляются с помощью 1.9:

$$\begin{gathered} C_0 = \log_e p, \\ C_m = a_m + \sum_{m-1}^{k=1} \frac{k}{m} C_k a_{m-k} \quad \text{при } 1 < m < p, \\ C_m = \sum_{m-1}^{k=m-p} \frac{k}{m} C_k a_{m-k} \qquad \text{при } m > p. \end{gathered} \tag{1.9}$$

1.4. Существующие наборы речевых данных

1.5. Классификаторы, используемые в SER

1.6. Постановка задачи

2 Конструкторский раздел

3 Технологический раздел

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. 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.
10. 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).
11. *Киселев В.*, *Давыдов А.*, *Ткаченя А.* Система определения эмоционального состояния диктора по голосу. 2012.
12. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge / B. Schuller, A. Batliner, S. Steidl, D. Seppi // Speech communication. 2011. Т. 53, № 9/10. С. 1062—1087.
13. Первичный анализ речевых сигналов. [Электронный ресурс], URL: https://alphacephei.com/ru/lecture1.pdf (дата обращения: 17.2.2023).

14. *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.

15. *Судьенкова А. В.* Обзор методов извлечения акустических признаков речи в задаче распознавания диктора // Сборник научных трудов Новосибирского государственного технического университета. 2019. № 3/4. С. 139—164.

ПРИЛОЖЕНИЕ А