УДК 004 **Информационные технологии**

Богданова Диана Радиковна, *кандидат технических наук, доцент,*
Уфимский государственный авиационный технический университет (г. Уфа)
Акушев Антон Табрисович, *магистрант,*
Уфимский государственный авиационный технический университет (г. Уфа)

РАСПОЗНАВАНИЕ ЭМОЦИЙ ПО РЕЧЕВОМУ СИГНАЛУ

Аннотация: В данной статье рассматриваются основные методы классификации эмоций по речевому сигналу, а также определена пошаговая модель распознавания эмоций на основе одного из рассматриваемых методов классификации.

Ключевые слова: Аффективные вычисления, речевой сигнал, классификация эмоций, распознавание эмоций, аффекты человека.

Abstract: This article discusses the main methods of classifying emotions based on a speech signal, and also defines a step-by-step model for recognizing emotions based on one of the classification methods under consideration.

Keywords: Affective computing, speech signal, emotion classification, emotion recognition, human affects.

Введение

Искусственный интеллект развивается семимильными шагами с каждый годом наращивая всё большее сообщество разработчиков в этой сфере. Одной из передовых технологий, в которой участвует искусственный интеллект, является распознавание эмоций. Оно применяется в маркетинге; сферах развлечений, например, для того, чтобы узнать реакцию зрителей на фильм; в сфере образования – с его помощью можно изучать настроение и внимание учеников во время занятий; так же эмоциональный ИИ может быть полезен в работе с

персоналом - он помогает определить состояние сотрудника, вовремя заметить его усталость или недовольство и эффективнее перераспределить задачи. Но как же именно ИИ определяет эмоции людей? Делает он это одним (или сразу несколькими) из следующих способов:

- Распознавание по изображению или видеопотоку (Визуальная информация).
- Распознавание по голосу (Аудио – Звуковая информация).
- Распознавание по жестам.
- Распознавание по вегетативным реакциями.

Чаще всего применяются именно распознавания по лицу или по голосу, во многом благодаря простоте доступа к анализируемым данных.

В различных источниках есть множество примеров распознавания эмоций на основе видеопотока данных или просто по изображениям, поэтому предлагаю затронуть чуть менее популярный способ распознавания – распознавание по аудиопотоку (По голосу).

Преобразование оригинального речевого сигнала

Голос часто отражает скрытые эмоции через высоту и тон, например, именно по интонациям голоса собаки могут понять настроение хозяина [1].

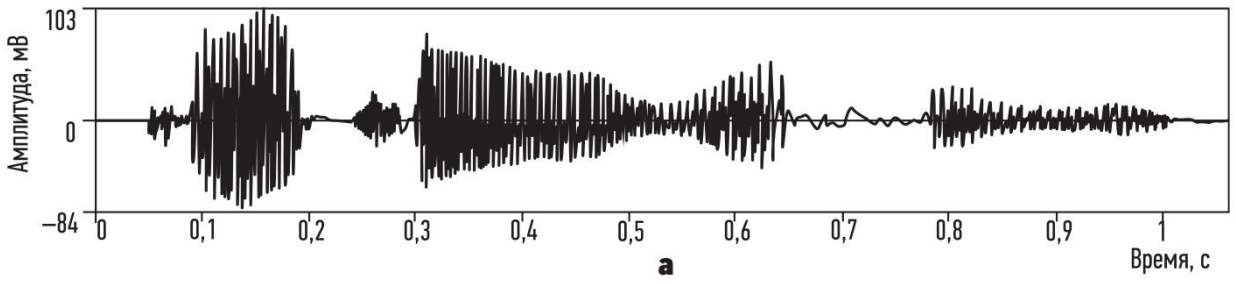

Рисунок 1 – Оригинальный речевой сигнал

Для классификации речевого сигнала необходимо преобразовать оригинальный речевой сигнал (Рисунок 1) в его спектрограмму (Рисунок 2).

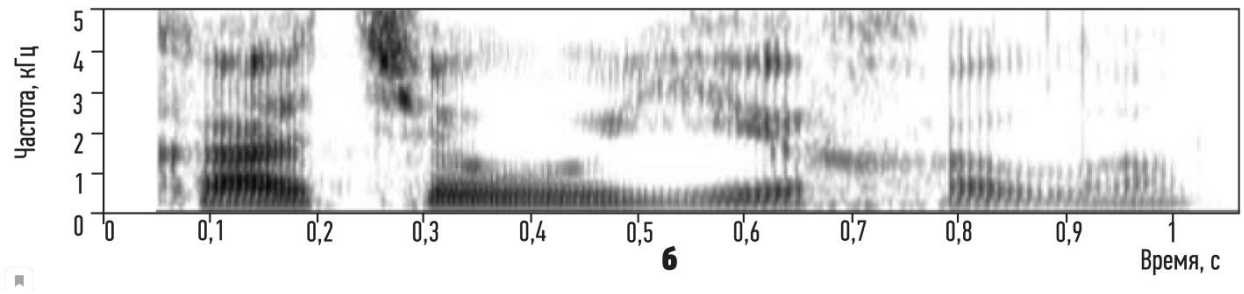

Рисунок 2 – Спектрограмма речевого сигнала

Спектрограмма речевого сигнала — это двухмерная диаграмма, вертикальная ось которой является осью частот, горизонтальная – осью времени, интенсивность же каждой точки на спектрограмме характеризует амплитуду речевого сигнала на определенной частоте в определенный момент времени. Она представляет из себя изображение, состоящее из оттенков серого. Для того чтобы определить временные границы сегментов, необходимо преобразовать спектрограмму в черно-белое изображение, т.е. определить такой порог, ниже которого точка будет считаться черной, а выше – белой [2]. Для этого могут быть использованы различные методы кластеризации.

Методы кластеризации речевого сигнала

Кластеризация – это процесс разделения исходной группы объектов на несколько подгрупп, называемых кластерами, таким образом, что объекты, находящиеся в одной подгруппе сильнее похожи друг на друга по тем или иным признакам, чем на объекты из других подгрупп. Кластеризация имеет применение во многих сферах: распознавание образов, машинное обучение, интеллектуальный анализ данных и так далее [3].

1. Распознавание с использованием машинного обучения (K-means)

Одним из самых простых и из-за этого фактора одним из самых часто используемых методов решения задачи является вычисление математического ожидания для частей кадра сигнала с последующим применением алгоритма K-means для классификации. K-means относится к методам с четкой кластеризацией: такие методы жестко ограничивают принадлежность элементов множества к одному кластеру. Он используется для распределения m

наблюдений по k кластерам таким образом, чтобы каждое из наблюдений принадлежало только одному кластеру, к центру которого оно наиболее приближено. В данном случае под наблюдениями понимаются точки на спектрограмме. Недостатком данного метода является невысокая точность распознавания эмоций [2].

2. Скрытые марковские модели

СММ играют важную роль в распознавании речи с 60-х годов [4] прошлого столетия. Их предназначение лежит в создании гибкой модели, соответствующей временным характеристикам речи, таким образом классифицируя мельчайшие вариации аудиосигнала. Естественно, СММ были одним из первых вариантов, которые можно было попробовать в SER (Speech Emotion Recognition). Некоторые исследователи пытались решить эту задачу, используя частотные представления, такие как MFCC и LPCC [5], а исследования были сосредоточены на низкоуровневых дескрипторах, таких как высота звука [6; 7].

3. Машина опорных векторов (SVM)

Машины опорных векторов (SVM) - оптимальный маржинальный классификатор в машинном обучении. Он также широко используется во многих исследованиях, связанных с распознаванием звуковых эмоций. SVM может иметь очень хорошие характеристики классификации по сравнению с другими классификаторами, особенно для ограниченных обучающих данных [8].

Классификация выполняется благодаря созданию N-мерной гиперплоскости, оптимально разделяющей данные по категориям. Основная идея этого метода лежит в использовании функций ядра, чтобы конвертировать исходные данные в многомерное пространство функций и последующего достижения оптимального разбиения по категориям в этом новом пространстве функций.

Функциональное ядро (RBF) используется на этапе обучения. Преимущество использования этого ядра заключается в том, что оно ограничивает обучающие данные указанными границами. RBF — это

нелинейное ядро, которое отображает образцы в пространство более высокой размерности с меньшими численными трудностями, чем полиномиальное ядро [7].

4. Искусственные нейронные сети

В машинном обучении и информатике многие проблемы невозможно преобразовать в простой алгоритм, который нужно решить. Решение этих проблем необходимо динамически адаптировать для каждой ситуации, в которой применяется алгоритм.

Адаптация - важная черта человеческого мозга, но в случае с компьютерами нет простого способа создание адаптивного кода. Наш мозг обладает способностью к обобщению, что помогает ему индуктивно рассуждать, что является самым первым шагом обучения. ИНС - отличное решение для адаптивного обучения. Они способны изучать сложные нелинейные отношения между входами и желаемыми выходами, эти системы сегодня широко используются почти во всех областях машинного обучения, и SER не является исключением [9].

В своем исследовании Shaw et al. [10] создали систему распознавания с использованием искусственных нейронных сетей для распознавания четырех классов эмоций: счастье, злость, грусть и нейтральные эмоции. Чтобы реализовать свою систему, они включают как просодические, так и спектральные характеристики для задачи классификации. Их сеть имеет входной слой, один скрытый слой и выход четырех классов. Результат их работы – более 80% общей точности для всех их классов.

Внедрение и обучение ИНС происходит быстрее, чем внедрение и обучение аналогичных методов на основе нейронных сетей, но однослойная ИНС не может решать очень сложные и нелинейные задачи, это граница, на которой ИНС останавливается. Таким образом, по сути, они могут быть быстрыми, но ограничены в своих возможностях [7].

5. Сверточные нейронные сети

Сверточные нейронные сети (CNN) — это особые типы нейронных сетей,

которые в своем скрытом слое имеют разные фильтры или области, которые реагируют на конкретную особенность входного сигнала. Их конструкция основана на исследовании Хьюбела и Визеля в 1968 году [11], в котором зрительная нервная кора представляет собой пространственно специализированную структуру, в которой каждая область реагирует на определенную характеристику входного сигнала. Одна положительная перспектива CNN — это способность изучать особенности из входных данных большого размера, однако они также узнают особенности от небольших вариаций и искажений внешнего вида, что приводит к необходимости большого объема памяти во время разработки. Следовательно, в CNN обычно существует слой свертки, за которым следует механизм понижающей дискретизации. Сверточный слой имеет различные банки фильтров, в которых их веса будут регулироваться посредством обратного распространения ошибки.

В своем исследовании Д. Бертеро и П. Фунг [12] представили сверточную нейронную сеть, способную обнаруживать эмоции гнева, радости и грусти с точностью более 66%. Они сравнили результаты своей обученной сети с базовой SVM на основе функций. Чтобы обучить и протестировать свой метод, они использовали корпус выступлений TED, помеченных студентами и созданными краудсорсингом. Они реализовали свою CNN с помощью инструментария Theano. Для сравнения была обучена линейная SVM с набором функций из эмоционального челленджа INTERSPEECH 2009.

Они сообщили, что их сеть CNN с временем отклика менее пары сотен миллисекунд способна обнаруживать классы злых, счастливых и грустных эмоций с точностью 66,1%. Они показали, что их нейронная активность сконцентрирована вокруг основных частот, наиболее соответствующих эмоциям [7].

Краткий анализ обзора

Если отбирать лучшие методы по параметру точности, то логично предположить, что лидирующие позиции будут занимать методы с использованием нейронных сетей. Именно поэтому самым релевантным

методом из рассмотренных является метод с применением искусственных нейронных сетей, благодаря быстрому процессу обучения и высокой точности распознавания.

Пошаговая модель распознавания эмоций по речевому сигналу

Работа типовой системы распознавания эмоций из речевого потока может быть описана следующей последовательностью шагов:

1. Получение на вход оригинального речевого (звукового) сигнала.

2. Преобразование входных данных в спектрограмму речевого сигнала в линейной или mel-шкале (Рисунок 3). Преобразование осуществляется с помощью алгоритма дискретного преобразования Фурье.

3. Сглаживание спектра сигнала. Может быть реализовано различными оконными фильтрами. Часто используемым фильтром является оконная функция Хэмминга.

4. Разбиение спектра сигнала на кадры.

Далее с полученными кадрами спектра оперируют как с обычными двумерными изображениями.

5. Вычисление значимых характеристик и признаков кадров.

6. Классификация эмоций на основе значимых характеристик кадра с использованием метода искусственных нейронных сетей.

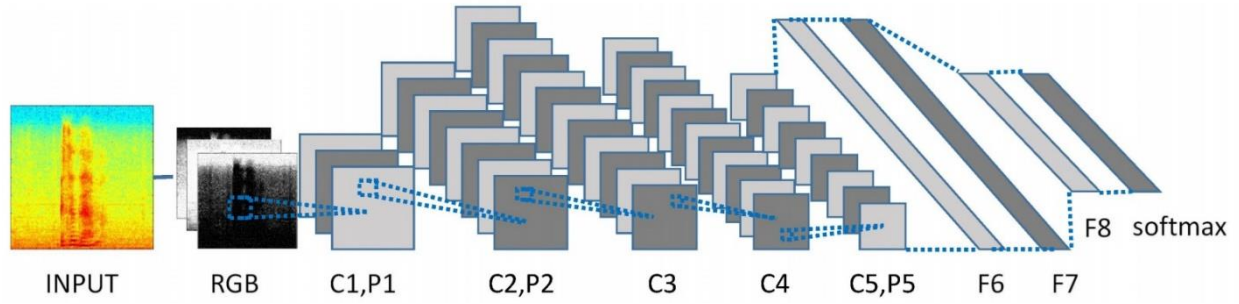

Рисунок 3 – Преобразование речевого сигнала в спектрограмму в линейной или mel-шкале[13]

Заключение

Сфера распознавания эмоций лишь начинает развиваться и в будущем можно будет автоматизировать множество процессов в различных сферах жизнедеятельности людей благодаря аффективным вычислениям, которые так или иначе зависят от эмоциональной составляющей людей.

В этой статье были проанализированы различные подходы для классификации эмоций по речевому сигналу, среди которых наилучшие результаты показывают методы с применением нейронных сетей, а также была изложена пошаговая модель классификации эмоций по голосу, ядром которой стала классификация с помощью метода искусственной нейронной сети.

Данная статья является отправной точкой для наших дальнейших исследований в области распознавания речевых эмоций, в которых нам предстоит использовать рассмотренные здесь методы классификации для реализации распознавания человеческих эмоций на реальных задачах.

Результаты исследований, приведенные в статье, получены в рамках выполнения грантов РФФИ 18-07-00193, 19-07-00709 и государственного задания № FEUE-2020-0007.

Библиографический список:

1. NewTechAudit «Распознавание эмоций с использованием нейронной сети», год: 2021, URL: https://vc.ru/dev/239027-raspoznavanie-emociy-s-ispolzovaniem-neyronnoy-seti.

2. Д.С. Канищев «АВТОМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ РЕЧИ МЕТОДАМИ КЛАСТЕРИЗАЦИИ И С ПРИМЕНЕНИЕМ МЕТОДА ОЦУ» - Вятский государственный университет, год: 2019, URL: https://cyberleninka.ru/article/n/avtomaticheskaya-segmentatsiya-rechi-metodami-klasterizatsii-i-s-primeneniem-metoda-otsu.

3. Задача кластеризации, URL: https://neerc.ifmo.ru/wiki/index.php?title=Кластеризация.

4. Rabiner, L.R. «A tutorial on hidden Markov models and selected applications in speech recognition». Proc. IEEE. Год: 1989. URL: https://ieeexplore.ieee.org/document/18626.

5. Nwe, T.L.; Foo, S.W.; De Silva, L.C. «Speech emotion recognition using hidden Markov models». Speech Commun. Год: 2003, URL: https://www.sciencedirect.com/science/article/abs/pii/S0167639303000992?via%3Dihub.

6. Nogueiras, A.; Moreno, A.; Bonafonte, A.; Mariño, J.B. Speech emotion recognition using hidden Markov models. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001.

7. Babak Joze Abbaschian, Daniel Sierra-Sosa, Adel Elmaghraby «Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models» - University of Louisville, год: 2021, URL: https://www.mdpi.com/1424-8220/21/4/1249/pdf.

8. Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub «Speech Emotion Recognition: Methods and Cases Study» Univerisity of Sousse, Tunisia, год: 2018, URL: https://www.researchgate.net/publication/322873355_Speech_Emotion_Recognition_Methods_and_Cases_Study.

9. Kriesel, D. Chapter 1: Introduction, Motivation and History. In A «Brief Introduction to Neural Networks» pp. 21–25. Available online: http://www.dkriesel.com (accessed on 2 February 2021).

10. Shaw, A.; Vardhan, R.K.; Saxena, S. Emotion Recognition and Classification in Speech using Artificial Neural Networks. Int. J. Comput. Appl. 2016, 145, 5–9.

11. Hubel, D.H.; Wiesel, T.N. «Receptive fields and functional architecture of monkey striate cortex» J. Physiol. Год: 1968, URL: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1968.sp008455.

12. Bertero, D.; Fung, P. «A first look into a Convolutional Neural Network for speech emotion detection». Год: 2017, URL: https://ieeexplore. ieee.org/document/7953131.

13. Калиновский И. (@kalisto21) «Введение в задачу распознавания эмоций» Опубликовано на интеллектуальном портале «Хабр» в 2018 году, URL: https://habr.com/ru/company/speechpro/blog/418151/.