Инженерный вестник Дона, №4 (2024)
ivdon.ru/ru/magazine/archive/n4y2024/9152

Реализация модели автоматического распознавания эмоций человека по речи

Д.А. Барышев, А.С. Зубанков, В.Л. Розалиев

Волгоградский государственный технический университет

Аннотация: Определение эмоций человека по речи является актуальной задачей на данный момент, потому как оно может быть применено в различных отраслях, таких, как экономика, медицина, маркетинг, безопасность и образование. В данной работе рассматривается распознавание человеческих эмоций именно по речи, т.к. речь является информативным показателем, который достаточно трудно подделать. В работе рассматривается нейросетевой подход к решению задачи. Была реализована рекуррентная нейронная сеть с памятью LSTM, а также собран собственный датасет, на котором модель обучалась. Датасет включает в себя речь русскоговорящих актеров, что повысит качество работы модели для русскоговорящих пользователей.

Ключевые слова: нейронная сеть, определение эмоций, речь, классификация, глубокое обучение, рекуррентная модель, LSTM.

В данной работе рассматривается нейросетевой подход для решения проблемы. Алгоритмичный подход также применим, однако он имеет ряд недостатков, из-за которых выбор пал именно на нейросетевой:

1. Трудоемкость работы – выполнить определение эмоций алгоритмично тяжелее и значительно дольше, чем реализовать модель нейронной сети[1];

2. Субъективность данных – мы не можем точно сказать, какие именно показатели и параметры речи относятся к той или иной человеческой эмоции. Определение четких границ будет однозначно субъективным, потому как часто различить конкретную эмоцию не представляется возможным, а также при определении эмоции разные люди могут дать разные варианты ответа;

3. Качество работы не будет выше, т.к. по сути придется вручную проделать ту же работу, которую выполняет модель автоматически. Здесь же следует отметить субъективность, которая нивелируется за счет большого датасета нейронной сети, а значит качество работы модели увеличивается [2,3].

Все решение поставленной задачи можно разбить на 2 большие части:

- обучение нейронной сети;

- определение эмоции в голосе человека с использованием обученной модели нейросети[4].

Под обучением нейронной сети подразумевается сбор и разметка (если необходимо) датасета и определение опциональных параметров модели.

Основной алгоритм работы происходит с уже обученной моделью. Т.е. в программе предусмотрена возможность запоминать обученное состояние модели и записывать его в файл.

Реализованная программа принимает на вход аудиофайл в формате «wav». Файл должен содержать записанную речь человека. Размер файла ограничен 100МБ. Допускается ввод входных данных через микрофон, а также загрузка wav-файла с помощью интерфейса пользователя. Если параметров командной строки нет, то программа должна брать входные данные из корневой папки.

Выходные данные представляют из себя текст, содержащий преобладающую эмоцию в речи.

Если пользователь записывает аудиозапись с помощью микрофона, то результирующий файл записывается в директорию с исполняемым файлом в файл с названием ‘test.wav’.

Программа может быть использована в рамках обучающего процесса, для тестирования работы нейронных сетей или в каких-либо других целях, в которых требуется автоматическое определение эмоций человека по речи.

Алгоритм работы программы подразумевает, что к моменту тестирования мы уже имеем обученную нейронную сеть. На вход поступает аудиофайл в формате wav. Это может быть, как путь к файлу на локальном

компьютере пользователя, так и аудиозапись, сделанная с помощью диктофона, в зависимости от выбора пользователя.

Полученная аудиозапись может содержать посторонние шумы, которые могут оказать влияние на итоговую работу программы[5]. Эти посторонние шумы необходимо удалить.

Далее мы получаем аудиозапись, очищенную от посторонних шумов. Однако мы не можем работать с аудиозаписью в сыром виде, т.к. объем данных слишком большой для обработки нейронной сетью. Поэтому, после очищения аудиозаписи, мы начинаем выделять признаки, которые мы будем подавать на вход нейронной сети[6,7]. К этим признакам относятся:

- Melspectogram (шкала Mel);
- Тоннетц;
- Спектральный контраст;
- Хромаграмма;
- MFCC.

Все эти признаки мы можем получить из аудиозаписи с помощью библиотеки librosa.

Конкатенируя каждый полученный признак, мы получаем данные, которые можем загрузить в обученную модель нейронной сети для определения эмоций.

Алгоритм работы программы можно разделить на 2 части – обучение модели нейронной сети и определение эмоций на уже обученной модели.

Алгоритм обучения модели представлен на рисунке 1.

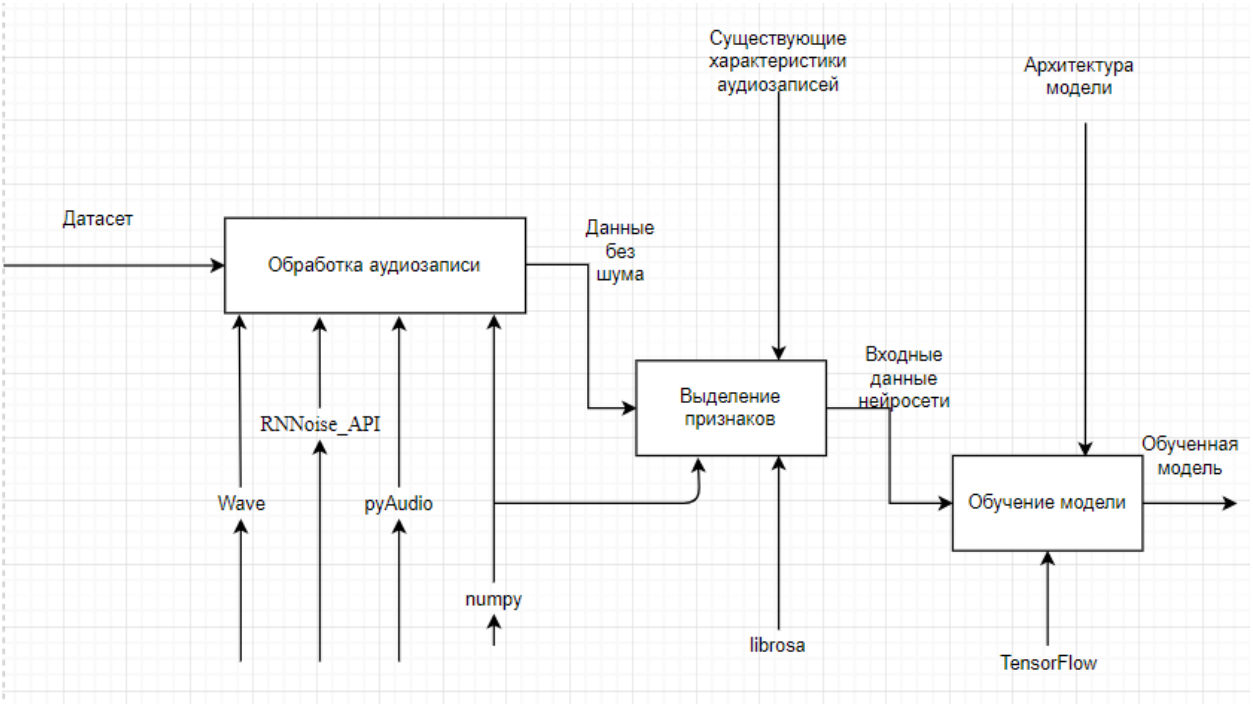

Рисунок 1. - IDEF0 диаграмма алгоритма обучения модели.

Одной из проблем для качественной реализации программы является наличие посторонних шумов во входящих аудиозаписях. Следует учитывать, что модель обучалась на датасете, созданном с помощью профессиональных актеров в профессиональных студиях звукозаписи, а значит, количество сторонних шумов и прочих неточностей сведено к минимуму, чего нельзя сказать об аудиозаписях, в которых определяет эмоцию уже обученная модель. Программа позволяет загрузить файл, в котором не проводилось предварительной ручной обработки шумов. Кроме того, программа позволяет записать входящую запись с помощью микрофона, что также увеличивает вероятность появления сторонних звуков. Для того, чтобы привести входящие записи к виду, в котором возможно определить эмоцию, либо же повысить качество определения эмоции, был использован алгоритм очистки шумов из аудиозаписи RNNoise_Wrapper. RNNoise_Wrapper представляет собой удобную обёртку над еще одной рекуррентной нейронной сетью, названной RNNoise, которая удаляет посторонние шумы.

Алгоритм работы программы с обученной моделью представлен на рисунке 2.

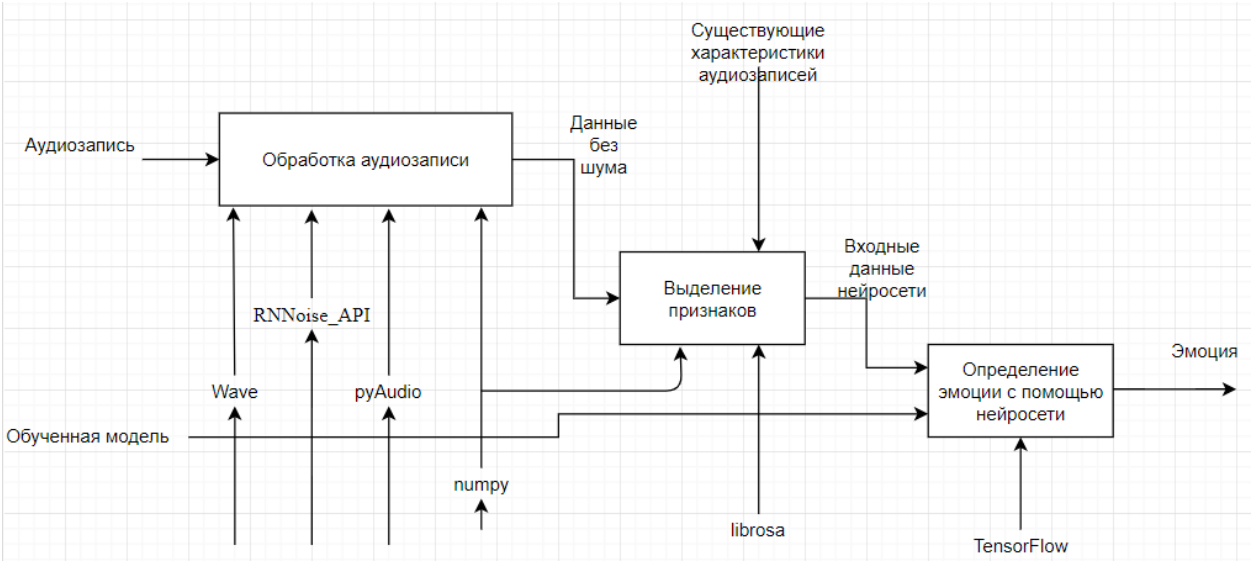

Рисунок 2. - IDEF0 диаграмма алгоритма работы программы.

Оптимизирован алгоритм обучения модели таким образом, чтобы обучение происходило один раз, запоминалось и более не требовалось переобучения для сокращения времени работы программы при последующих запусках. Эта оптимизация реализована с помощью механизмов, представленных в библиотеке TensorFlow языка Python.

Для обучения нейронной сети были взяты наборы данных на английском языке из открытых источников TESS и RAVDESS. С целью улучшения точности работы модели для русскоязычных пользователей был собран набор данных из русскоязычных открытых источников. Как правило, примеры брались из записей экзаменов студентов актерских ВУЗов по дисциплине «Эмоциональная речь», а также тренингов профессиональных актеров с произношением текстов с определенной эмоцией. Следует отметить, что такой метод сбора данных делает оценку эмоций в аудиозаписях наиболее объективной, исключая ошибку человека, собирающего датасет. Все эмоции размечены самими актерами. В датасете

размечены такие эмоции, как гнев, радость, удивление, страх, грусть, отвращение и нейтралитет.

Получено всего:

- 52 аудиозаписи с эмоцией «гнев»;
- 50 аудиозаписей с эмоцией «радость»;
- 56 аудиозаписей с эмоцией «удивление»;
- 58 аудиозаписей с эмоцией «страх»;
- 57 аудиозаписей с эмоцией «грусть»;
- 52 аудиозаписи с эмоцией «отвращение»;
- 59 аудиозаписей с эмоцией «нейтралитет» [8].

После анализа моделей нейронных сетей, решающих аналогичные задачи, была выбрана рекуррентная нейронная сеть с долгосрочной памятью LSTM, т.к. она наиболее подходит для задач классификации последовательностей [9,10]. Опытным путем была найдена оптимальная архитектура, выдавшая точность, равную 91.56%. Состав архитектуры нейронной сети был следующий: 2 слоя RNN со 128 нейронами на каждом слое, а также 2 Dense-слоя с размерностью 128 нейронов на слое. Был выбран Batch size = 64 и количество эпох, равное 1000. Для задачи классификации используется функция потерь – категориальная кросс-энтропия [11]. Такая функция потерь наиболее оптимальна для многоклассовой классификации, а в рамках нашей задачи - классов 8 по числу определяемых эмоций.

Архитектура модели нейронной сети представлена на рисунке 3.

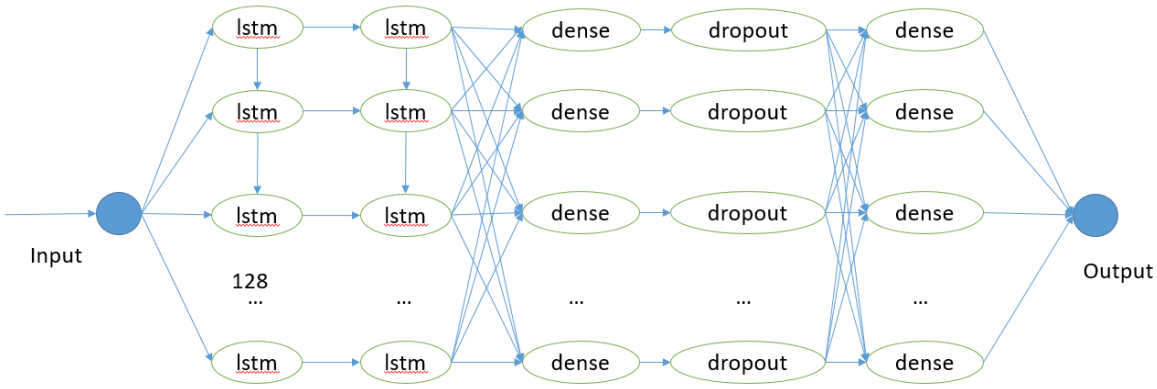

Рисунок 3. - Архитектура модели нейронной сети

Таким образом, в данной статье представлен разработанный метод решения задачи. Метод представляет собой нейросетевую модель LSTM на вход которой поступает стек фич, таких, как melspectogram (шкала Mel), тоннетц, спектральный контраст, хромаграмма и MFCC, а на выходе пользователь получает эмоцию, которая преобладает в аудиозаписи. Также описан алгоритм работы программы, который, в свою очередь, разделен на 2 основные части – обучение модели нейронной сети и определение эмоций человека по аудиозаписи с использованием обученной нейронной сети [12,13].

В статье описана архитектура полученной модели, рассмотрено, что поступает на вход и на выход, какие используются слои и их размеры, а также остальные параметры модели. Описаны использованные датасеты, в том числе собранный вручную датасет русскоговорящих актеров.

Литература

1. Вартанян Г.А., Петров Е.С. Эмоции и поведение. - Л.: Наука, 2009. - 144 с.

2. Вудвортс Р. Выражение эмоций // Экспериментальная психология. - М., 2000. - 798 с.

3. Марьев. А.А. Метод интерпретации результатов измерений

параметров речевого сигнала в задачах диагностики психоэмоционального состояния человека по его речи // Инженерный вестник Дона, 2011, №4. URL: ivdon.ru/ru/magazine/archive/n4y2011/538.

4. Каллан, Р. Основные концепции нейронных сетей. Пер. с англ. – М.: Издательский дом «Вильямс». 2001. – 288 с.

5. Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. "Comparison of different implementations of MFCC." Journal of Computer Science and Technology 16.6 (2001): pp. 582-589.

6. Сидоров К.В., Ребрун И.А., Кожевников Д.Д., Соботницкий И.С. Диагностика психофизиологического и эмоционального состояния человека-оператора // Инженерный вестник Дона, 2012, №4 (часть 2). URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1480.

7. Васильев И.А., Поплужный В.Л. Тихомиров О.К. Эмоции и мышление. - М., 2010. - 288 с.

8. Экман Пол. Психология эмоций. - 2-е изд. / Пер. с англ. — СПб. Питер, 2010. – 27с.

9. Ayadi El M., Kamel M. S., Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases // Pattern Recognition. – 2011. – V. 44. – №. 3. – pp. 572-587.

10. Mower Provost E., “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow”. -. IEEE, 2013. - IEEE ICASSP no. 17, pp. 4-17, 2013.

11. Грибачев, В.П. Настоящее и будущее нейронных сетей. Компоненты и технологии, №5, 2006. – С. 28-32.

12. Bengio Y., Courville A., Vincent P., “Representation learning: A review and new perspectives”. - IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

13. Ithaya Rani P., Muneeswaran K. Facial Emotion Recognition Based on Eye and Mouth Regions // International Journal of Pattern Recognition and Artificial Intelligence. – 2016. – V. 30. – №. 07. – pp. 79-81.

References

1. Vartanyan G.A., Petrov E.S. Emocii i povedenie [Emotions and behavior]. L.: Nauka, 2009. 144 p.

2. Vudvorts R. Eksperimental'naya psihologiya., M., 2000, 798 p.

3. Mar'ev A.A. Inzhenernyj vestnik Dona, 2011, №4. URL: ivdon.ru/ru/magazine/archive/n4y2011/538.

4. Kallan, R. Osnovnye koncepcii nejronnyh setej [Basic concepts of neural networks]. Per. s angl. M.: Izdatel'skij dom «Vil'yams». 2001. 288 p.

5. Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. Journal of Computer Science and Technology 16.6 (2001): pp. 582-589.

6. Sidorov K.V., Rebrun I.A., Kozhevnikov D.D., Sobotnickij I.S. Inzhenernyj vestnik Dona, 2012, №4 (chast' 2). URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1480.

7. Vasil'ev I.A., Popluzhnyj V.L. Tihomirov O.K. Emocii i myshlenie [Emotions and thinking]. M., 2010, 288p.

8. Ekman Pol. Psihologiya emocij [The psychology of emotions]. 2-e izd. Per. s angl. SPb.: Piter, 2010, 27p.

9. Ayadi El M., Kamel M. S., Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 2011, V. 44, № 3, pp. 572-587.

10. Mower Provost E., IEEE, 2013. IEEE ICASSP no. 17, pp. 4-17, 2013.

11. Gribachev, V.P. Nastoyashchee i budushchee nejronnyh setej [Present and future of neural networks]. Komponenty i tekhnologii, №5, 2006. pp. 28-32.

12. Bengio Y., Courville A., Vincent P., IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

13. Ithaya Rani P., Muneeswaran K. Facial Emotion Recognition Based on Eye and Mouth Regions. International Journal of Pattern Recognition and Artificial Intelligence. 2016. V. 30. №. 07. pp. 79-81.

Дата поступления: 1.03.2024

Дата публикации: 11.04.2024