Нейросетевой подход к определению эмоций человека по речи

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

Волгоградский государственные технический университет

Аннотация: С точки зрения практической ценности определение эмоций в голосе человека может применяться в разных областях, связанных как с передачей аудиосообщений, так и в общении онлайн: к таким сферам относятся медицина, безопасность, экономика, образование и прочие. В качестве яркого примера можно представить оценку качества работы операторов call-центров, а также предлагаемых ими услуг/товара. Так, наличие сигналов о том, что клиент испытывает отрицательные эмоции, например, гнев, может говорить о возможных проблемах у операторов. В данной работе будет рассматриваться нейросетевой подход для автоматического определения эмоций человека по его речи.
Ключевые слова: нейронная сеть, определение эмоций, речь, классификация, глубокое обучение, свёрточная модель.

Работы по автоматическому определению эмоций проводятся уже давно. Существует несколько способов определения эмоций: по визуальным признакам - мимика, походка, речь (фонетически или контекстно); с помощью более глубокого анализа, например, электрокардиограммы. В данной работе мы будем рассматривать определение эмоций с помощью фонетической составляющей речи человека, т.е. по его голосу без привязки к смыслу.

Голос сообщает окружающим о текущем состоянии человека: о его переживаниях, отношении к фактам, самочувствии, а нередко — о темпераменте и чертах характера [1]. Более всего о внутреннем психоэмоциональном состоянии человека может сообщить анализ его связной речи: как в ней расставлены логические ударения, как быстро произносятся слова, как конструируются фразы, какие имеются отклонения от нормы. Эти реакции обычно плохо поддаются контролю, даже если их пытаются маскировать. Потому они весьма информативны [2].

При решении задач определения эмоций в голосе человека можно столкнуться с целым перечнем проблем [3]. Основными являются:

- Отсутствие размеченного датасета на нужном языке. К сожалению, в

открытом доступе отсутствуют датасеты на русском языке. Это связано с тем, что для составления объективно корректного датасета необходима команда актеров разных возрастов, рас, национальностей и пола. Так, если брать датасет англоговорящих людей для определения эмоций людей из азиатских стран, то результативность будет довольно мала. Это связано с менталитетом, семантическими и фонетическими особенностями языка (в некоторых языках одно и то же слово может иметь разное значение в зависимости от интонации, с которой его произносят). Также было бы неправильным решением определять эмоции в голосе ребенка с помощью нейросети, обученной на низком голосе взрослого человека. Для создания качественного набора входных данных требуется как можно больше треков с аудиозаписями (измеряемо в тысячах). Чем больше эмоций мы хотим рассмотреть, тем больше будет и датасет. К тому же, чтобы соблюсти объективность, на каждый пол, расу, возраст и т.д. требуется взять несколько разных актеров, т.к. понятие эмоции в голосе может отличаться в зависимости от субъективных взглядов человека. Все это увеличивает финансовые затраты на его создание.

- Наличие посторонних шумов в аудиозаписи. Эта проблема решается технически, однако даже современные методы ее решения неидеальны. Убрав часть посторонних звуков на записи, мы рискуем убрать лишнее, а также оставить часть того, что следовало бы убрать, что повлечет за собой снижение эффективности итогового решения. Все это осложняется наличием одновременно нескольких голосов на записи.

- Субъективность эмоций. Часто люди сами не могут сойтись во мнениях какая эмоция звучит в голосе. Усугубляется это еще тем, что в зависимости от ситуации одни и те же признаки эмоций в голосе могут интерпретироваться по-разному [4].

Учитывая перечень существующих критериев, задача определения эмоций по голосу человека является труднодостижимой. Более того, многие характеристики, по которым можно было бы разделять эмоции - пересекаются для разных эмоций, что также усложняют поставленную задачу.

Эмоция – это психическое состояние, связанное с нервной системой и вызванное химическими реакциями в организме, которое, как правило, является отражением мыслей, чувств и поведенческих реакций человека [5].

Для определения эмоций человека требуется некая классификация, которая поможет строго отличать одну эмоцию от другой. В качестве такой классификации мы будем использовать популярную градацию эмоций по Полу Экману [6], однако немного изменим ее, добавив еще одну эмоцию – нейтральную, что требуется для реализации системы автоматического распознавания эмоций по голосу человека[7,8].

Таким образом, получим перечень возможных определяемых эмоций человека:

- нейтралитет (спокойствие);
- страх;
- грусть;
- злость;
- отвращение;
- презрение;
- удивление;
- радость.

Были проанализированы существующие решения в данной предметной области. Рассматривая только определение эмоций по голосу человека и исключая логическую составляющую его речи, а также мимику, мы получили следующий перечень аналогов, которые представлены в таблице

№1:

Таблица №1

Анализ существующих аналогов

Система	Страна	API	Кол-во эмоций	Какие эмоции распознает	Критерии (по чему определяет)
Emotion AI	Япония	Да	6	радость, гнев, спокойствие, печаль, удивление, смущение	Паралингвистика, тон, громкость, скорость
Beyond Verbal	Израиль	Да	400 оттенков	Гнев, одиночество, счастье, волнение	Интонация, громкость голоса
Smart Logger II	Россия	Нет	2	Норма, не норма	Интонация, мелодичность, скорость, ритмичность, громкость речи, пол, возраст
Cogito	США	Нет	2	Беспокойство, гнев	Тон собеседника, паузы между фразами, скорость речи, изменения в голосе

Проанализировав источники, можно сделать вывод, что на данный момент не существует универсального решения, охватывающего несколько эмоций и оценивающих их по достаточному количеству характеристик.

Взяв лучшее из всех решений – количество эмоций (в нашей работе их 8), признаки сравнения эмоций, а также сделав решение доступным, т.е. не требующим дополнительного дорогостоящего оборудования или мощных серверов для запуска, мы предложили свое решение в виде нейронной сети, определяющей эмоции по речи.

Для решения задачи была выбрана рекуррентная модель нейронной сети. Такой выбор обусловлен тем, что рекуррентные нейронные сети –

лучше других моделей справляются с задачей классификации последовательностей [9,10]. Решение задачи с помощью алгоритмического подхода было сразу исключено ввиду долгого времени выполнения программы, а также отсутствия объективно точных критериев разделения эмоций. Даже вручную проанализировав датасет, на котором обучается текущая модель, невозможно сделать однозначный вывод для точной границы признаков отношения классов эмоций. Поэтому в качестве решения поставленной задачи был выбран именно нейросетевой подход.

Из-за сложности определения эмоции по какой-то одной конкретной характеристике, в работе используется 5 разных характеристик:

1. Шкала Mel - это перцептивная шкала тонов, которые слушатели оценивают как равные по расстоянию друг от друга. Контрольная точка между этой шкалой и нормальным измерением частоты определяется путем присвоения тону 1000 Гц высоты восприятия 1000 мел, что на 40 дБ выше порога слушателя. Выше примерно 500 Гц, слушатели оценивают все более большие интервалы, чтобы получить равные приращения высоты тона.

2. Высота тона — насколько высок или низок звук. Это зависит от частоты, более высокий шаг - это высокая частота.

3. Частота — скорость вибрации звука, измеряет волновые циклы в секунду.

4. Цветность (Хроматограмма)— представление звука, в котором спектр проецируется на 12 ячеек, представляющих 12 различных полутонов (или цветность). Вычисляется путем суммирования спектра значений логарифмической частоты по октавам. Есть теории, что каждая эмоция имеет свой цвет. Несмотря на то, что данные из разных источников разнятся, можно выделить некую моду и, ориентируясь на ее значение, вывести эмоции по звуку (из хронограммы).

5. MFCC — Mel Частотные Кепстральные коэффициенты: голос зависит от формы голосового тракта, включая язык, зубы и т.д. Представление кратковременного спектра мощности звука, по сути, представление голосового тракта [11].

Для обучения нейронной сети выбран датасет Crema-D – эмоциональный мультимодальный актерский набор данных. Он содержит в себе 7442 аудиозаписи 91 актера с различным этническим происхождением. В датасете, а также во всей работе рассматриваются 8 базовых человеческих эмоций – радость, грусть, злость, удивление, отвращение, презрение, страх, нейтралитет (отсутствие ярко выраженной эмоции). Также добавлены 48 кастомных аудиозаписей на русском языке, по 8 на каждую эмоцию. Эти записи взяты из музыкальных, театральных и кинопредставлений на русском языке для обогащения датасета.

В качестве модели нейронной сети выбрана LSTM – рекуррентная нейронная сеть с возможностью запоминания состояния предыдущих слоев, а также полносвязные слои. В качестве алгоритма оптимизации был выбран Adam. Опытным путем была найдена оптимальная архитектура, выдавшая точность равную 86.74%. Для этого происходило обучение на 4 слоях – 2 RNN размерность 128 нейронов на слое и 2 Dense с такой же размерностью – 128 нейронов на слое. Был выбран Batch size = 64 и количество эпох, равное 1000. Batch – это количество итераций для одной эпохи. Такая структура необходима для лучшей работы нейронной сети, т.к. мы не можем отправить в нее сразу все записи. Для задачи классификации используется функция потерь – категориальная кросс-энтропия. Такая Loss функция выбрана из-за того, что количество классов больше 2 [12].

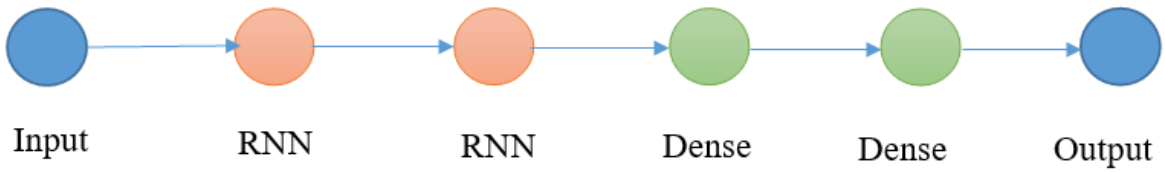

Рис. 1. – Визуализация слоев нейронной сети

После обучения нейронной сети главной проблемой становится качество поступающего аудиопотока. Если в датасете все аудиозаписи были чистыми и размеченными, то входные тестируемые записи могут содержать массу посторонних звуков, мешающих корректной работе программы. Решение этой проблемы - в реализации алгоритма шумоподавления. Для решения этой проблемы в программе используется алгоритм RNNoise_Wrapper, который является удобной оберткой над нейронной сетью RNNoise, которая позволяет удалять шум из аудиопотока в реальном времени.

Таким образом, получаем следующий алгоритм работы программы (рис. 2):

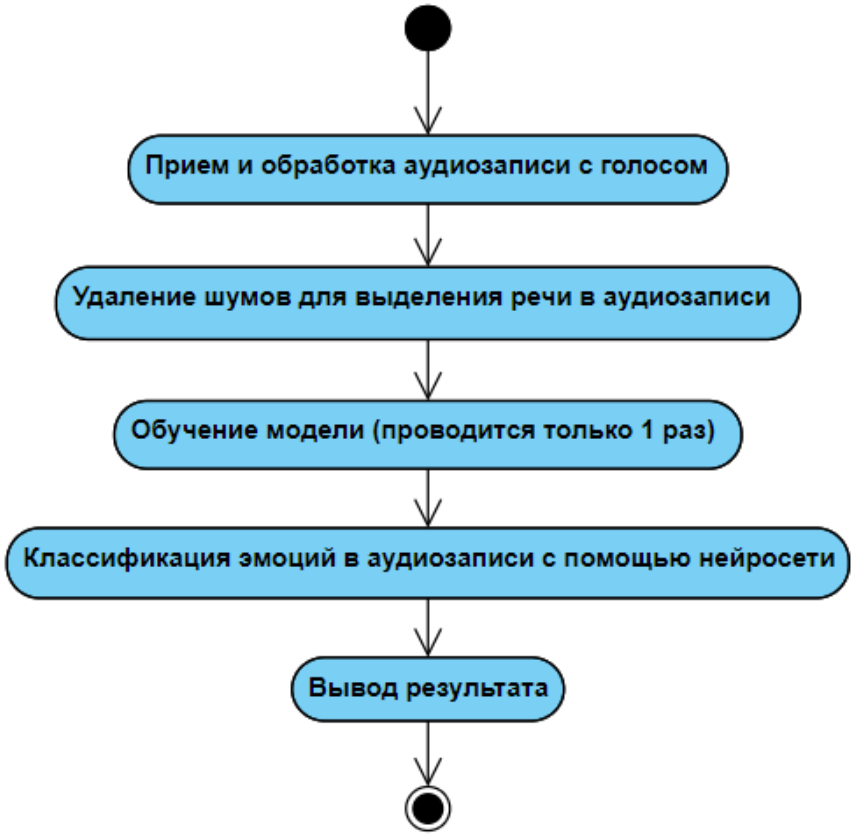

Рис. 2. – Алгоритм работы программы

Модель обучается на локальном ПК только один раз, т.к. TensorFlow позволяет сохранить обученное состояние нейросети, что позволяет последующие запуски производить без обучения.

Для тестирования нейронной сети выбраны следующие метрики:

- Accuracy - точность (правильно классифицированных примеров ко всем примерам) как метрика оценки качества обучения сети (1) [13].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (1)$$

где T – true, F- false, P – positive, N – negative.

- Precision и Recall – точность и полнота.

Precision (2) можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющихся положительными, а recall (3) показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм [13].

$$Precision = \frac{TP}{TP+FP}, \quad (2)$$

$$Recall = \frac{TP}{TP+FN} (3), \quad (3)$$

Результат для 3 эмоций (счастье, грусть, нейтралитет/спокойствие) следующий: Accuracy = 86,74%, Precision = 82,17%, Recall = 89,01%.

Результат для 8 эмоций по Экману - радость, грусть, злость, удивление, отвращение, презрение, страх, нейтралитет - следующий: Accuracy = 71,81%, Precision = 79,30%, Recall = 74,98%.

Опираясь на полученные значения точности нейронной сети, можно сделать вывод, что нейронная сеть работает корректно и показывает достаточно высокий результат для данного количества классов эмоций.

Итого, результатом текущей работы является программа, определяющая эмоции человека по его голосу, с помощью свёрточной

нейронной сети. Программа выдает в результате ту эмоцию, которая преобладала в аудиозаписи дольше всего.

В дальнейшем планируется увеличить точность нейронной сети с помощью увеличения датасета, а соответственно, и количества эпох обучения, а также добавления Dropout-слоев в архитектуру модели. Также планируется сделать визуальный пользовательский интерфейс для более удобной работы с программой.

Литература

1. Вартанян Г.А., Петров Е.С. Эмоции и поведение. Ленинград: Наука, 1989. 144 с.

2. Вудвортс Р. Выражение эмоций // Экспериментальная психология. Москва, 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. 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.

5. Васильев И.А., Поплужный В.Л. Тихомиров О.К. Эмоции и мышление. Москва, 2010. 288 с.

6. Экман Пол. Психология эмоций. - 2-е издание. Пер. с англ. Санкт-Петербург: Питер, 2010. 27 с.

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

8. Mower Provost E., “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow”. International Conference on Acoustics, Speech and Signal Processing. Vancouver, British Columbia, Canada. May, 2013.

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

10. Каллан Р. Основные концепции нейронных сетей. Пер. с англ. Москва: Издательский дом «Вильямс». 2001. 288 с.

11. 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.

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, 2013, pp. 1798–1828.

13. Ithaya Rani P., Muneeswaran K. Facial Emotion Recognition Based on Eye and Mouth Regions //I nternational Journal of Pattern Recognition and Artificial Intelligence. 2016. Т. 30. №. 07.

References

1. Vartanyan G.A., Petrov E.S. Emocii i povedenie [Emotions and behavior]. Leningrad: Nauka, 1989. 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. 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.

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

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

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

8. E. Mower Provost, “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow”. International Conference on Acoustics, Speech and Signal Processing. Vancouver, British Columbia, Canada. May, 2013.

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

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

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

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

13. Ithaya Rani P., Muneeswaran K. International Journal of Pattern Recognition and Artificial Intelligence. 2016. T. 30. №. 07.