**Исследование возможностей классификации эмоций в невербальном речевом поведении на различных наборах данных**

**Аннотация**

**Ключевые слова:** эмоциональные вычисления, распознавание эмоций, визуализация многомерных данных

**1. Введение**

Распознавание эмоций человека является важной и актуальной задачей ввиду активного развития и расширения систем человеко-машинного взаимодействия, цифровых систем коммуникации и обеспечения совместной деятельности людей: систем электронной коммерции, человеко-машинных интерфейсов, социо-киберфизических систем, систем Интернета вещей, видеоконференцсвязи и т.д. При использовании таких систем могут происходить нежелательные ситуации, связанные с конфликтами и деструктивным поведением людей. В таких ситуациях особенно важной становится оценка эмоционального состояния пользователей, которое является важнейшим предиктором поведения человека.

Эмоции человека являются «сложными психическими процессами и состояниями, связанными с инстинктами, потребностями, мотивами и отражающих в форме непосредственного переживания (удовлетворения, страха, радости и т.д.) значимость действующих на индивида явлений и ситуаций для осуществления его жизнедеятельности. Сопровождая практически любые проявления активности субъекта, эмоции служат одним из главных механизмов внутренней регуляции психической деятельности и поведения, направленных на удовлетворение актуальных потребностей» [Вилюнас В. К. Эмоции [элетронный ресурс] // Большой психологический словарь/под общ. ред. Б.Г. Мещерякова, В.П. Зинченко <https://psychological.slovaronline.com/2078-EMOTSII> (дата обращения хх.хх.2020)]. В структуру эмоций человека входят следующие компоненты: импрессивная (субъективное переживание эмоции), когнитивная компонента (связанные с эмоцией представления, мысли, установки, осознаваемый опыт), физиологическая (биохимические изменения, изменения в работе внутренних органов, изменения в работе центральной и периферической нервной системы), поведенческая или моторная компонента, которая, в том числе служит для экспрессивной функции эмоций. Ввиду того, что физиологическая и поведенческая компоненты могут быть внешне измеряемыми и наблюдаемыми, они представляют особый интерес для исследований в области автоматического распознавания эмоций человека. При этом, анализ поведенческой экспрессии эмоций человека можно анализировать на расстоянии без непосредственного участия человека, в то время как методы, основанные на физиологической компоненте, требуют непосредственного участия человека в изменениях, высоких временных и трудовых затрат на измерения со стороны испытуемых и экспериментаторов, а также наличия дорогостоящего оборудования для проведения измерений. Экспрессивная компонента эмоций может быть представлена в различных модальностях: мимические выражения, поза и двигательная активность тела субъекта, невербальное речевое поведение, вербальное речевое поведение. Наряду с другими модальностями, невербальное речевое поведение может быть использовано для автоматического опосредованного распознавания эмоций человека. При этом, распознавание на основе этой модальности становится особенно актуальным в случае недостатка или отсутствия данных других модальностей, а также в моделях, методах и системах многомодального распознавания эмоций.

Современные модели, методы и системы распознавания эмоций человека основываются, в основном, на концепции базовых эмоций – выделении некоторого набора элементарных эмоций, не сводимых к другим [Ortony, Andrew & Turner, Terence. (1990). What's Basic About Basic Emotions?. Psychological review. 97. 315-31. 10.1037/0033-295X.97.3.315.]. Данная концепция принимается в качестве основания для категоризации данных для выполнения дальнейшей классификации системами автоматического распознавания эмоций. Некоторыми исследователями невербальное речевое поведение считается основанием для выделения базовых эмоций [Scherer K. R. Vocal affect expression: A review and a model for future research //Psychological bulletin. – 1986. – Т. 99. – №. 2. – С. 143.][Scherer K. R., Johnstone T., Klasmeyer G. Vocal expression of emotion //Handbook of affective sciences. – 2003. – С. 433-456.][Banse R., Scherer K. R. Acoustic profiles in vocal emotion expression //Journal of personality and social psychology. – 1996. – Т. 70. – №. 3. – С. 614.]. Наборы данных [S. R. Livingstone и F. A. Russo, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English», PLOS ONE, т. 13, вып. 5, с. e0196391, май 2018.][ "F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in 9th European Conference on Speech Communication and Technology (Interspeech’2005 - Eurospeech), Lisbon, Portugal, September 2005, pp. 1517–1520."], использующие разметку, основанную на концепции базовых эмоций, однако, могут содержать в себе различный набор классов эмоций, использующихся для категоризации данных. Это может быть обусловлено несколькими причинами: применение различных оснований для выделения базовых эмоций, не достаточная представленность той или иной эмоции в наборах, не согласованность при разметке данных экспертами и т.д. Не однородность наборов данных ставит важнейшую научную задачу сравнительного исследования различных алгоритмов классификации на большом наборе различных наборов данных, содержащих в себе невербальные проявления эмоций человека в речи.

Стоит также особо отметить, что в некоторых специфических задачах на первый план выходят вопросы переразметки известных наборов данных на другие категории. Так, в задаче выявления агрессивного поведения людей, на передний план выходит выявление негативных эмоций, которые обычно сопровождают проявления агрессии [Комалова Л.Р.. "Перцептивно-слуховой профиль (образ) агрессора" Вестник Московского государственного лингвистического университета. Гуманитарные науки, no. 7 (746), 2016, pp. 116-126.][Комалова Л.Р.. "Сопоставление слухового и зрительного видов восприятия агрессивного речевого поведения" Вестник Московского государственного лингвистического университета. Гуманитарные науки, no. 15 (754), 2016, pp. 114-128.]. Такая переразметка может быть выполнена простым отнесениям экземпляров того или иного класса эмоций к более общей категории – негативные эмоции и остальные эмоции, к которым относят позитивные и нейтральные эмоции. Так, к негативным эмоциям обычно относят гнев, печаль, отвращение, страх, а к остальным – счастье, волнение, скука, удивление, нейтральную эмоцию. Не смотря на всю свою простоту и очевидность, данному подходу уделяется недостаточно внимания в современной научной литературе.

Наряду с обозначенными выше проблемами, стоит также уделить внимание проблеме оценки данных на предмет их группировки в том или ином признаковом пространстве. Такая задача может быть выполнена при помощи различных процедур проекции данных, представленных в признаковом пространстве большой размерности, в признаковое пространство малой размерности с последующей визуализацией данных в этом пространстве. Это позволяет выполнить предварительную качественную оценку данных на предмет их применимости в задаче классификации.

Целью данной работы является выполнение визуализации и качественной оценке данных, содержащих невербальные речевые проявления эмоций человека для различных наборов данных, а также выполнение сравнительного исследования алгоритмов классификации, обученных на этих наборах.При этом, и визуализация, и классификация выполняются на основании двух способов категоризации данных, обозначенных выше, – использование классов, представленных в наборе денных по умолчанию, а также разбиение на два класса эмоций – негативные и остальные.

Статья организована следующим образом: обзор литературы и методов решения задачи приведен в разделе 2. В разделе 3, мы описываем выбранные нами подходы к визуализации и классификации. В разделе 4 представлены описания и результаты экспериментов, их интерпретация и обсуждение – в разделе 5. В разделе 6 мы подводим итоги работы и намечаем вектор дальнейшего исследования.

**2. Обзор литературы**

Разработка модели машинного обучения для классификации человеческих эмоций по голосу является нетривиальной задачей, поскольку голосовой сигнал содержит в себе много информации, как напрямую относящейся к передаче текущего эмоционального состояния человека, так и вовсе не относящейся к эмоциям. Поэтому решение задачи автоматического распознавания эмоций в первую очередь подразумевает определение некоторого релевантного набора признаков, извлекаемого из звуковой записи человеческой речи. Признаки, извлекаемые из звукового сигнала, делятся на низкоуровневые дескрипторы (low-level-descriptors, LLD) и функциональные признаки. Низкоуровневые дескрипторы включают в себя просодические (высота тона, громкость, энергия, тембр, продолжительность пауз и др.) и спектральные (фундаментальная частота, часоты основных формант, мел-кепстральные частотные коэффициенты (MFCC), кепстральные коэффициенты линейного предсказания (LPCС) и др.) характеристики, а так же их производные по времени. Функциональные признаки включают в себя статистические показатели низкоуровневых дескрипторов (минимум, максимум, различные процентили, zero-crossing-rate).

Результатом работы авторов [Eyben, Florian & Wöllmer, Martin & Schuller, Björn. (2010). openSMILE -- The Munich Versatile and Fast Open-Source Audio Feature Extractor. MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. 1459-1462. 10.1145/1873951.1874246.][Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor”, In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. doi:10.1145/2502081.2502224] стал инструмент OpenSMILE, позволяющий извлекать широкий спектр параметров звукового сигнала, а также применять различные функции к этим параметрам. Благодаря этому, появились наборы параметров, которые могут быть легко извлечены с помощью этого инструмента. К этим наборам относятся как очень большие, как, например, стандартные наборы признаков для конференций INTERSPEECH, содержащие более 5000 элементов [Schuller, Björn & Steidl, Stefan & Batliner, Anton & Vinciarelli, Alessandro & Scherer, Klaus & Ringeval, Fabien & Chetouani, Mohamed & Weninger, Felix & Eyben, Florian & Marchi, Erik & Mortillaro, Marcello & Salamin, Hugues & Polychroniou, Anna & Valente, Fabio & Kim, Samuel. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 148-152.], так и попытки унифицировать признаковое пространство для задач аффективных вычислений [F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," in IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 1 April-June 2016, doi: 10.1109/TAFFC.2015.2457417].

Наиболее распространенным подходом в современных исследованиях является извлечение комбинации вышеописанных параметров из акустического сигнала с последующей конкатенацией их в признаковые векторы, которые в дальнейшем используются для обучения классификатора, выбор которого также является важным шагом в решении задачи. К наиболее популярным в задаче распознавания эмоций по голосу классификаторам относятся: машина опорных векторов (SVM), скрытые марковские модели (HMM), гауссовская смешанная модель (GMM), алгоритм k ближайших соседей (k-NN), различные архитектуры глубоких нейронных сетей. Так, Lin и др [Lin, J.C.; Wu, C.H.; Wei, W.L.: Semi-coupled hidden Markov model with state-based alignment strategy for audio-visual emotion recognition, in Proc. Affective Computing and Intelligent Interaction (ACII), 2011, 185–194.] продемонстрировали высокую степень релевантности просодических характеристик в задаче распознавания эмоций. Их набор признаков состоял из высоты тона, энергии, частот пяти основных формант F1-F5 и был использован для обучения скрытой марковской модели. Точность классификации составила 67,8%. Данный результат, однако, был получен на собственном наборе данных, содержащем четыре эмоциональные категории. В исследовании Eyben [Eyben, F.; Petridis, S.; Schuller, B.; Pantic, M.: Audiovisual vocal outburst classification in noisy acoustic conditions, in ICASSP, 2012, 5097–5100.] 9 акустических низкоуровневых дескрипторов были использованы для обучения глубокой рекуррентной нейронной сети с долгой краткосрочной памятью (LSTM-RNN). На наборе данных TUM AVIC для классификации одного из четырех классов получены результаты Unweighted Average Recall 67,6%, однако их алгоритм направлен на выявление сложных аффективных состояний (смех, уверенность, неуверенность, остальное), нежели базовых эмоций. Laithia и др. [Lalitha, S., & Tripathi, S. (2016). Emotion detection using perceptual based speech features. 2016 IEEE Annual India Conference (INDICON). doi:10.1109/indicon.2016.7839028], напротив, извлекали из аудиосигнала исключительно спектральные характеристики: мел-кепстральные частотные коэффициенты (MFCC), кепстральные коэффициенты линейного предсказания (LPCС), Perceptual Linear Predictive Cepstrum (PLPC), Mel Frequency Perceptual Linear Predictive Cepstrum (MFPLPC), фундаментальная частота f0, амплитуда, фаза, а также их статистические функции, получив 64-размерный вектор, который был использован для обучения глубокой нейронной сети. Получен результат в 85,7% точности распознавания одной из семи эмоций на наборе данных Emo-DB. Стоит отметить, что в их работе отсутствует описание использованной архитектуры глубокой нейронной сети. Metallinou и др. [Metallinou, A.; Lee, S.; Narayanan, S.: Audio-visual emotion recognition using Gaussian mixture models for face and voice in Proc. Int. Symp. Multimedia, 2008, 250–257.] использовали 39-размерный признаковый вектор на основе MFCC для обучения смешанной гауссовской модели, получив среднюю точность распознавания на наборе данных IEMOCAP 54.34%, однако не использовали этот набор данных целиком, ограничившись образцами, представленными лишь четырьмя базовыми эмоциями. Наконец, Petrushin [Petrushin VA (2000) Emotion recognition in speech signal: experimental study, development, and application. In: Proceedings of ICSLP, pp 222–225] в своей работе извлек из сигнала просодические и спектральные низкоуровневые дескрипторы и их статистические функции, получив 43 параметра, из которых с помощью алгоритма RELIEF-F [Kononenko, I. Estimating attributes: Analysis and extension of RELIEF. In L. De Raedt and F. Bergadano (eds.) Proc. European Conf. On Machine Learning. 171-182, 1994.] было выбрано 14 наиболее важных, которые и вошли в итоговый признаковый вектор. На предмет результатов классификации был проведен сравнительный анализ алгоритма k-NN, пятислойного перцептрона и ансамбля нейронных сетей. На собственном наборе данных, содержащем 5 эмоциональных категорий, этими моделями была достигнута средняя взвешенная точность классификации 55%, 65% и 70% соответственно.

Одним из state-of-the-art подходов на данный момент является построение спектрограмм, что сводит задачу к классификации изображений. Satt и др. [Satt, Aharon, Shai Rozenberg and Ron Hoory. “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms.” INTERSPEECH (2017).] предложили метод глубокого обучения 5-слойной сверточной нейронной сети на предварительно полученных спектрограммах очищенного от шума звукового сигнала, получив точность классификации на наборе данных IEMOCAP 66%, однако использовали из этого набора данных лишь 4 базовых эмоциональных класса.

Наиболее прогрессивными и эффективными на данный момент являются так называемые end-to-end подходы, которые работают непосредственно с дискретизированным аудиосигналом в формате WAV и в которых предобработка, извлечение релевантных признаков и классификация объединены в единый «черный ящик». Для таких подходов обычно используются сверточные глубокие нейронные сети в комбинации с другими архитектурами. Trigeorgis и др. [Trigeorgis, George & Ringeval, Fabien & Brueckner, Raymond & Marchi, Erik & Nicolaou, Mihalis & Schuller, Björn & Zafeiriou, Stefanos. (2016). Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Recurrent Network. 10.13140/RG.2.1.3842.7283.] представили архитектуру глубокой нейронной сети, состоящей из сверточных и LSTM слоев. Для сравнения эффективности их модели со стандартным эвристическим подходом, был так же извлечен набор признаков, который был использован для обучения SVM и глубокой нейронной сети долгой краткосрочной памяти. Предложенный авторами метод превзошел по качеству классификации обе базовые модели, показав результат на наборе данных RECOLA 68%. ElShaer и др.[ElShaer, Mohamed & Wisdom, Scott & Mishra, Taniya. (2019). Transfer Learning From Sound Representations For Anger Detection in Speech.] описывают метод распознавания эмоций с помощью трансферного обучения пятислойной глубокой сверточной нейронной сети SoundNet[Aytar, Yusuf & Vondrick, Carl & Torralba, Antonio. (2016). SoundNet: Learning Sound Representations from Unlabeled Video.]. Авторы рассматривают задачу бинарной классификации "гнев"-"не гнев". Их архитектура - SoundNet, у которой первые два сверточных слоя "замораживают", последующие три инициализируются случайно. Получившаяся нейросеть дообучается на наборе данных IEMOCAP. Авторами заявлено улучшение качества классификации в сравнении с обучением той же нейронной сети со всеми пятью слоями, инициализированными случайно), а также значительное превосходство в генерализации у трансферно обученной модели над обученной "с нуля", однако отсутствует сравнение их модели с какой-либо другой, что ставит под вопрос ее применимость.

Для визуализации многомерных данных наиболее часто используются алгоритмы снижения размерности, среди которых анализ главных компонент [Ivosev, G., Burton, L. and Bonner, R., 2008. Dimensionality reduction and visualization in principal component analysis. *Analytical chemistry*, *80*(13), pp.4933-4944.], линейный дискриминантный анализ [Wang, H., Ding, C. and Huang, H., 2010, September. Multi-label linear discriminant analysis. In *European conference on computer vision* (pp. 126-139). Springer, Berlin, Heidelberg.], алгоритм t-distributed stochastic neighbor embedding (t-SNE)[ L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.].

**3. Методология**

3.1. Извлекаемые признаки

Первой и важной проблемой в разработке модели машинного обучения для нашей задачи является выбор подходящего признакового пространства. На текущий момент, между исследователями не существует единого консенсуса относительно релевантности тех или иных параметров акустического сигнала применительно к задаче распознавания эмоциональных состояний по голосу, однако существуют попытки к стандартизации параметров, используемых в эмоциональных вычислениях, такие, как набор признаков Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," in IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 1 April-June 2016, doi: 10.1109/TAFFC.2015.2457417]. Этот набор содержит 88 параметров звукового сигнала, среди которых: статистические функции от основной частоты и амплитуды, мел-кепстральные коэффициенты (MFCC) 1-4, приближенная оценка количества слов в секунду. Эти параметры наиболее полно отражают основные акустические признаки эмоций [Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31(1), 39–58 (2009)]. Также положительной особенностью данного набора параметров является независимость размерности итогового признакового вектора от длительности звукового сигнала. Признаки извлекаются с помощью инструмента OpenSMILE.

3.2. Алгоритмы классификации

В качестве сравниваемых классификаторов выбраны: алгоритм k ближайших соседей (k-NN) и машина опорных векторов (SVM).

Алгоритм k ближайших соседей определяет класс объекта путем определения класса большинства объектов из числа k ближайших соседей объекта . Для оценки расстояния между объектами в kNN-алгоритме могут использоваться различные меры расстояния, такие как евклидова мера, манхэттенская мера, косинусная мера и др [Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K. and Kerdprasop, N., 2015, March. An empirical study of distance metrics for k-nearest neighbor algorithm. In Proceedings of the 3rd international conference on industrial application engineering (pp. 280-285).].

Основная идея метода опорных векторов, также известного в литературе как Машина опорных векторов (SVM) заключается в отображении векторов признаков в пространство более высокой размерности, после чего в последнем строится разделяющая гиперплоскость. При этом все векторы, расположенные с одной "стороны" гиперплоскости, относятся к одному классу, а расположенные с другой - ко второму [Вапник, В. Н., Червоненкис, А. Я. Теория распознавания образов. — М.: Наука, 1974. — 416 с.][ Vladimir N. Vapnik, . "The Nature of Statistical Learning Theory." (1999).].

3.3. Предобработка данных

Выбранные алгоритмы классификации основаны на оценке евклидового расстояния в признаковом пространстве, поэтому масштабирование данных влияет на качество классификации как для k-NN[], так и для SVM[Tax, D.M. and Duin, R.P., 2000. Feature scaling in support vector data descriptions. *Learning from Imbalanced Datasets*, pp.25-30.]. В связи с этим, с целью приведения данных к единому масштабу значений все векторы признаков были подвергнуты z-нормализации. Этот метод представляет из себя масштабирование данных на основе среднего по выборке значения и стандартного отклонения.

3.4. Визуализация многомерных данных

Чтобы получить представление о распределении данных в признаковом пространстве, использован алгоритм уменьшения размерности t-SNE. Он является одним из самых распространенных методов визуализации многомерных данных. Суть данного метода заключается в проекции данных больших размерностей в пространства, обладающие меньшей размерностью, с сохранением расстояний между экземплярами данных, близко расположенных друг относительно друга. Это позволяет группировать однородные данные в пространствах малой размерности. Перед применением алгоритма также производится z-нормализация выборки.. Полученные таким образом визуальные представления отражают взаимное расположение данных различных классов и позволяют предварительно оценивать обрабатываемые данные на возможность применения их для дальнейшей классификации.

**4. Эксперименты**

4.1. Использованные наборы данных

IEMOCAP (Interactive emotional dyadic motion capture database) [ Busso, Carlos & Bulut, Murtaza & Lee, Chi-Chun & Kazemzadeh, Abe & Mower Provost, Emily & Kim, Samuel & Chang, Jeannette & Lee, Sungbok & Narayanan, Shrikanth. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 42. 335-359. 10.1007/s10579-008-9076-6.]

CREMA-D (Crowd-sourced emotional multimodal actors dataset) – это аудиовизуальный корпус, опубликованный в 2014 году. ["H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-sourced emotional multimodal actors dataset,” IEEE Transactions on Affective Computing, 2014."] Он состоит из видеозаписей эмоциональных проявлений на английском языке. В записи принимали участие 91 актеров (48 мужчин и 43 женщины), которым было предложено прочесть 12 предложений с одной из шести эмоциональных окрасок: радость, печаль, гнев, страх, отвращение, нейтральная. Таким образом было получено 7442 клипа, аудиодорожка представлена в формате wav с частотой дискретизации 48 кГц. В разметке и валидации корпуса через краудсорсинг было задействовано 2443 человека, точность распознавания эмоциональной окраски по звуковой модальности составила 40,9%.

Emo-DB[F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in 9th European Conference on Speech Communication and Technology (Interspeech’2005 - Eurospeech), Lisbon, Portugal, September 2005, pp. 1517–1520.]

RAVDESS[S. R. Livingstone и F. A. Russo, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English», PLOS ONE, т. 13, вып. 5, с. e0196391, май 2018.]

SAVEE[Jackson, Philip & ul haq, Sana. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database]

TESS[M. K. Pichora-Fuller и K. Dupuis, «Toronto emotional speech set (TESS)». Scholars Portal Dataverse, 2020]

Имеющиеся в нашем распоряжении наборы данных см. в Таблице 1. Здесь и далее приняты следующие сокращения для базовых эмоций: ANG - Anger (гнев); HAP - Happiness (радость); SAD - Sadness (печаль); NEU - Neutral (нейтральная эмоция); DIS - Disgust (отвращение); FEA - Fear (страх); BOR - Boredom (скука); SUR - Surprise (удивление); EXC - Excitement (возбуждение); FRU - Frustration (негодование); CAL - Calm (спокойствие).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Название | Год | Язык | Дикторы | Тип произносимого | N записей | Формат аудио | Базовая дискретная разметка |
| CREMA-D | 2014 | Англ. | 48 мужчин и 43 женщины, возраст от 20 до 74 лет, средний возраст - 36лет | 12 коротких предложений | 7442 | wav, 48 кГц, 2 канала | ANG, HAP, SAD, NEU, DIS, FEA |
| IEMOCAP | 2007 | Англ. | 5 мужчин и 5 женщин | Диалоги по сценарию и спонтанные импровизации | 7304 | wav, 48 кГц, 2 канала | ANG, HAP, SAD, NEU, EXC, FRU |
| Emo-DB | 2005 | Нем. | 5 мужчин и 5 женщин | 5 коротких и 5 длинных предложений | 535 | wav, 16кГц, 1 канал | ANG, HAP, SAD, NEU, DIS, FEA, BOR |
| RAVDESS | 2018 | Англ. | 12 мужчин и 12 женщин, от 21 до 33 лет, средний возраст - 26 лет | 2 коротких предложения | 1440 | wav, 48кГц, 2 канала | ANG, HAP, SAD, NEU, DIS, FEA, SUR, CAL |
| SAVEE | 2008 | Англ. | 4 мужчины, от 27 до 31 года | 120 предложений | 480 | wav, 44.1кГц, 2 канала | ANG, HAP, SAD, NEU, DIS, FEA, SUR |
| TESS | 2020 | Англ. | 2 женщины возрастом 26 и 64 года | Фраза 'Say the word X', где Х - одно из 200 односложных слов | 2800 | wav, 48кГц, 2 канала | ANG, HAP, SAD, NEU, DIS, FEA, SUR |

Таблица 1. Наборы данных, подлежащие исследованию.

Задача распознавания эмоций по голосу осложнена отсутствием единой методики составления размеченных наборов данных. Наборы данных, доступные для исследовательских целей, значительно отличаются друг от друга качеством записи, количеством дикторов и представленными базовыми дискретными эмоциональными категориями. В связи с этим, приняты решения:

* 1. Для каждого набора данных провести многоклассовую классификацию. Сравнить результаты, сопоставить их с визуализациями.
  2. Для каждого набора данных провести бинарную классификацию негативной эмоциональной окраски по стратегии «один против остальных». Для этого добавить в каждый из наборов данных вторую разметку: NEG (Negative) – Негативная эмоция, REST – остальные эмоциональные состояния. Категории ANG, SAD, FRU, DIS, FEA отнести к негативным эмоциям, HAP, EXC, NEU, BOR, SUR, CAL – к остальным; Проанализировать результаты бинарной классификации, сопоставить с результатами t-SNE.
  3. Путем слияния всех наборов данных на английском языке, получить обобщенный набор данных English Assembly, содержащий 19462 образца и произвести попытку как мультиклассовой, так и бинарной классификации. При этом, для мультиклассовой классификации принято решение ограничиться образцами, помеченными шестью самыми распространенными базовыми эмоциями согласно Ekman [Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion. Chichester: Wiley.]: ANG, HAP, DIS, FEA, NEU, SAD. В случае же бинарной классификации, используются все образцы.

4.2. Визуализация распределения данных в признаковом пространстве.

Мы используем библиотеку *ScikitLearn* [Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.], в которой имеется имплементация алгоритма t-SNE. Чтобы не перегружать визуализации и не усложнять работу алгоритма, количество подаваемых на вход алгоритма образцов данных было сокращено до случайной выборки, состоящей из 2000 (для объединенного набора данных English Assembly – из 5000) образцов с сохранением пропорций распределения классов эмоций. Использованные нами параметры подобраны исходя из соображений, описанных в [Poličar P.G., Stražar M., Zupan B. (2019) Embedding to Reference t-SNE Space Addresses Batch Effects in Single-Cell Classification. In: Kralj Novak P., Šmuc T., Džeroski S. (eds) Discovery Science. DS 2019. Lecture Notes in Computer Science, vol 11828. Springer, Cham]: *n\_components=2*, *n\_iter=5000*; *learning\_rate=0.1*; *early\_exaggeration=2*. Параметр perplexity варьировался от 5 до 75, таким образом, было получено по 70 визуализаций для каждого набора данных, из которых было выбрано по одной наиболее удачной.

Визуализации результатов t-SNE на исходных наборах данных при оригинальной разметке представлены на рисунке 1.

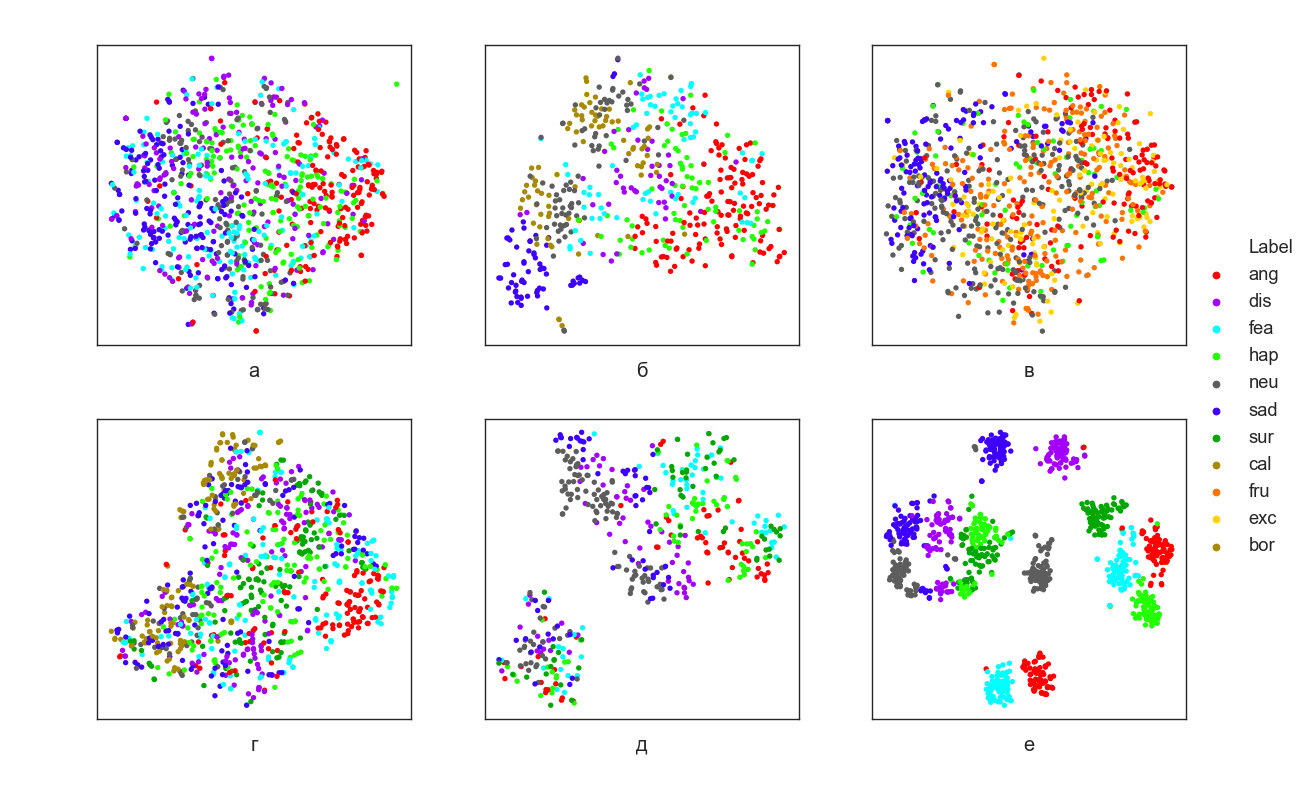


Рисунок 1. Визуализации результатов t-SNE на наборах данных при оригинальной разметке: а - CREMA-D, б - Emo-DB, в - IEMOCAP, г - RAVDESS, д - SAVEE, е - TESS

* CREMA-D (Рисунок 1, а), IEMOCAP (Рисунок 1, в), RAVDESS (Рисунок 1, г): одна группа данных, некоторые классы (ANG, SAD) имеют большую плотность распределения в определенных зонах этой группы, в то время как такие классы, как FEA, NEU распределены по всей области более равномерно.
* SAVEE (Рисунок 1, д): две группы данных, в обеих группах локализуются представители нескольких классов, однако внутри этих групп можно выделить области, наиболее представленные экземплярами одного из классов.
* Emo-DB(Рисунок 1, б): данные группируются в соответствии с классами, однако четких границ между этими группами нет, и различные классы распределены с различной плотностью.
* TESS(Рисунок 1, е): наиболее ярко выраженная структура среди всех полученных визуализаций, данные разбиты на плотные скопления с четкими границами, причем каждая эмоциональная категория представлена двумя группами. Можно утверждать, что данные сгруппированы не только по представленным классам базовых эмоций, но и по актерам, участвовавшим в записи.

Визуализации результатов t-SNE на исходных наборах данных при бинарной разметке представлены на рисунке 2.

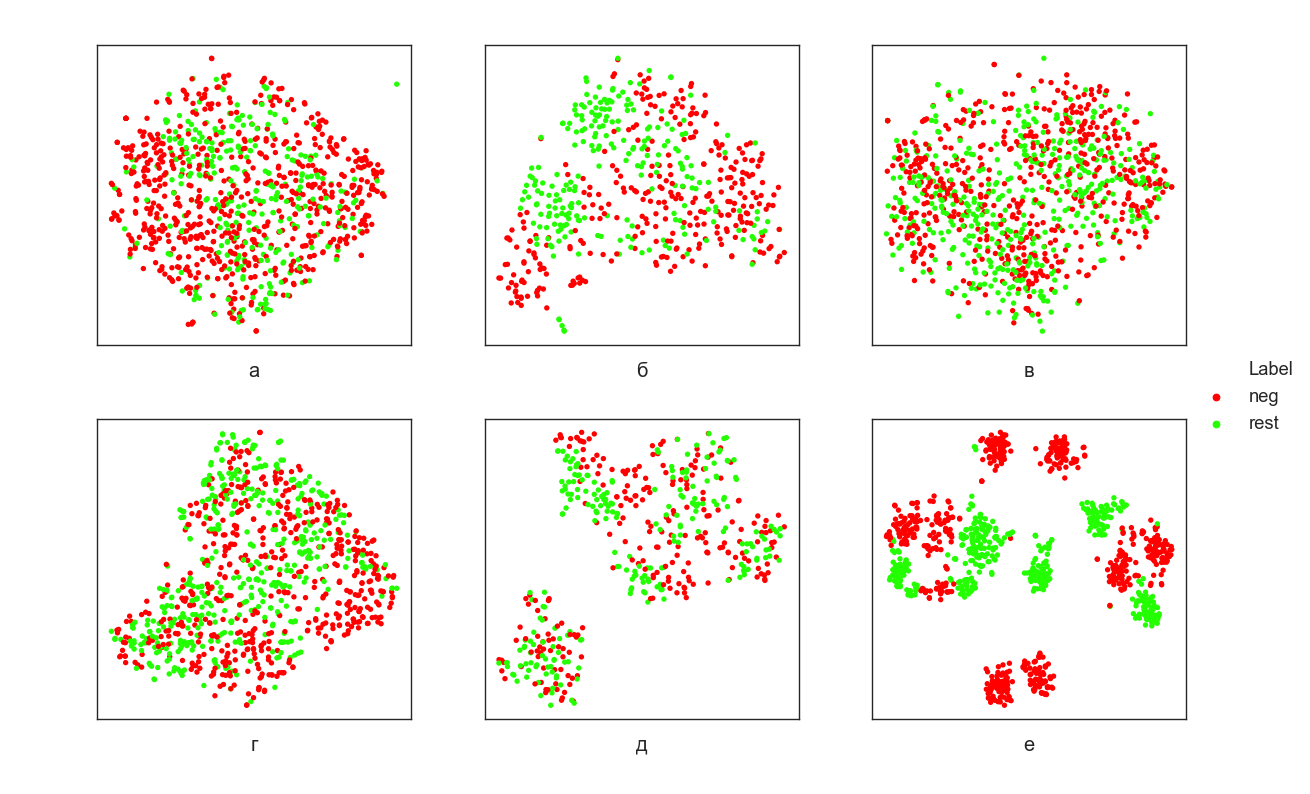


Рисунок 2. Визуализации результатов t-SNE на наборах данных при бинарной разметке: а - CREMA-D, б - Emo-DB, в - IEMOCAP, г - RAVDESS, д - SAVEE, е - TESS

* IEMOCAP (Рисунок 2, в), SAVEE (Рисунок 2, д): данные обоих классов распределены равномерно по всему пространству, нет яркой выраженности одного класса по сравнению с другим в той или иной области.
* RAVDESS (Рисунок 2, г), CREMA-D (Рисунок 2, а): существуют области пространства с преобладанием представленности одного класса над другим, однако четких границ между классами не наблюдается.
* TESS (Рисунок 2, е), Emo-DB (Рисунок 2, б): явно наблюдаем области пространства, представленные только одним из двух классов, эти области можно визуально разграничить.

Визуализации результатов t-SNE на объединенном наборе данных English Assembly при многоклассовой и бинарной разметке представлены на рисунке 3.

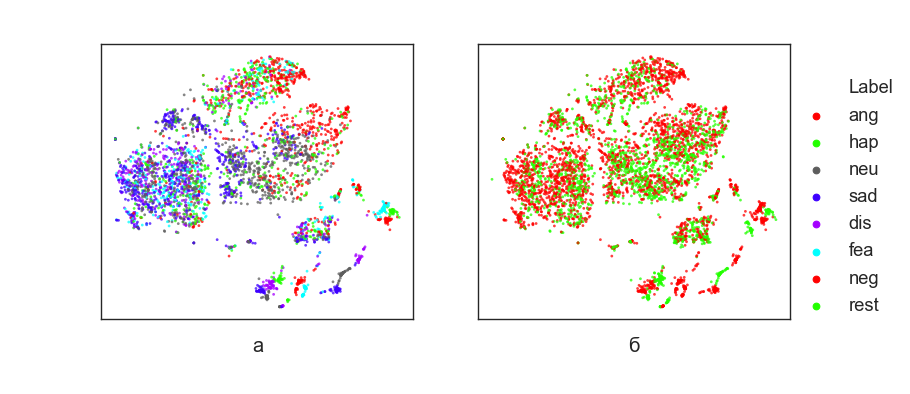


Рисунок 3. Визуализации результатов t-SNE на объединенном наборе данных English Assembly. a - многоклассовая разметка, б - бинарная разметка

Видно, что как в случае многоклассовой разметки (Рисунок 3, а), так и в случае бинарной разметки (Рисунок 3, б) данные сгруппированы в пространстве не по классам.

4.3. Классификация.

Для классификации также используется библиотека *ScikitLearn*, имеющая в своем составе классы, отвечающие за реализацию классификационных алгоритмов k-NN и SVM: *KNeighborsClassifier()* и *SVC()* соответственно. Для экспериментов нами был использован перебор параметров моделей с оптимизацией точности классификации. В случае алгоритма k-NN, параметр *k*, отвечающий за количество учитываемых ближайших соседей объекта, менялся в диапазоне от 2 до 75, остальные параметры принимали значения по умолчанию. В случае SVM, параметры варьировались по следующему набору: *kernel=linear*, *C=[0.001, 0.10, 0.1, 10, 25, 50, 100]*. Данные были предварительно разбиты на обучающую и тестовую выборку в пропорции 7:3, с сохранением пропорций распределения эмоциональных классов. В качестве метрики качества классификации выбраны мера точности  и мера полноты . где *TP* (True Positive rate), *FP* (False Positive rate), *FN* (False Negative rate) – соответственно количества истинно-положительных, ложно-положительных и ложно-отрицательных предсказаний модели. Таким образом, интуитивно,  демонстрирует способность алгоритма обнаруживать определенный класс, а  — способность отличать этот класс от других классов.

Полученные результаты представлены в таблице(ах?)

**5. Интерпретация и обсуждение результатов.**

По результатам классификации видно, что в подавляющем большинстве случаев, у алгоритма k ближайших соседей в случае мультиклассовой классификации не возникает проблем с распознаванием гнева, также достаточно уверенно распознается печаль и нейтральная эмоциональная окраска. Это видно также на визуализациях t-SNE, где образцы этих эмоций занимают строго определенную область пространства. В случае бинарной классификации негативной эмоциональной окраски по стратегии «один против остальных», можно наблюдать явный перевес качества классификации в сторону определения истинно «негативного» класса и большой ложноположительный показатель, что свидетельствует о некоторой степени переобученности модели, устранив которую, можно получить лучшие результаты.

На наборе данных TESS получены выделяющиеся результаты как классификации, так и визуализации, однако, это связано с особенностями этого набора данных – чрезмерное модулирование актерами эмоций, произнесение одной и той же фразы, всего два актера. В общем случае, лучшие результаты получены для тех наборов данных, в которых: а) Разделение и модуляция эмоциональных переживаний сильнее выражена (более яркая игра актеров); б) Произносятся одинаковые и короткие фразы; в) Меньшее количество актеров.

На некоторых наборах данных (SAVEE, TESS) T-SNE группирует данные по отличному от эмоциональной окраски признаку. Это могут быть кластеры для говорящих разного пола, или для разных актеров. Применение разделения по полам или использование идентификации говорящего может улучшить качество классификации.

**6. Выводы и дальнейшие планы.**

По итогу работы можно сделать вывод, что возможность и результаты классификации эмоций по речи методом k ближайших соседей во многом зависят от особенностей обучающего набора данных, а также от метода предобработки. В наши дальнейшие планы входит расширение списка используемых наборов данных (в том числе, на русском языке), методов классификации и предобработки в целях улучшения качества классификации. Также планируется применять рассмотренные методы в задачах мультимодального распознавания эмоций.

**7. Литература.**