

УДК 004.896

Взаимосвязь эмоциональной окраски речи с ее паралингвистическими признаками и параметрами текстовой информации

Киселев Виктор Валерьевич¹Васильев Александр Викторович² (научный руководитель)¹ МФТИ, Москва, Россия² ВолГУ, Волгоград, Россия

Цель исследования — изучение взаимосвязей эмоциональной окраски речи с ее паралингвистическими признаками и параметрами текстовой информации, разработка интерпретируемой модели предсказания эмоций. В ходе выполнения работы был проведен анализ взаимосвязи извлеченных высокоуровневых признаков между собой, с классами целевой переменной, обучена интерпретируемая модель на наиболее полезных параметрах, оценено ее качество, проведено сравнение с аналогичными моделями, определена важность признаков лучи

Ключевые слова: машинное обучение, распознавание эмоций, обработка аудио, важность признаков, интерпретируемость

Введение. Распознавание эмоций — одна из ключевых задач анализа речи. Актуальность работы подтверждается тем, что на производстве важно отслеживать эмоциональное состояние работников, чтобы предотвращать конфликты, депрессивные настроения. С каждым годом увеличивается распространение голосовых ассистентов, в текущий момент они широко распространены в домашних умных устройствах, происходят первые шаги их внедрения на заводы. Определение эмоции пользователя позволит учитывать его настроение при составлении ответов в диалоге с ним.

Постановка задачи. Дан набор аудиозаписей речи X , каждая запись которого представляет собой последовательность переменной длины из элементов, обозначающих амплитуду звукового сигнала. Дан фиксированный набор классов эмоций $Y = \{Y_1, Y_2, \dots\}$. Тренировочный датасет D состоит из пар $y_i \in Y, x_i \in X: \{(y_1, x_1), \dots (y_n, x_n)\}$. Итоговая задача — минимизировать функцию потерь на обучающем множестве:

$$\min_{\theta} L(\theta) = \frac{1}{|D|} \sum_{(y_i, x_i) \in D} l(y_i, f_{\theta}(x_i)), \quad (1)$$

где θ — параметры модели; f_{θ} — модель классификации; l — функция потерь.

Обзор существующих подходов. Стандартной практикой решения рассматриваемой задачи является обучение нейросетей. Например, в работе [1] для предсказания метки эмоции анализируются не только значения сигнала непосредственно, но и изображения спектрограмм, которые подаются на вход

сверточной нейронной сети с архитектурой VGG16 (рис. 1), обученной на наборе данных Imagenet.

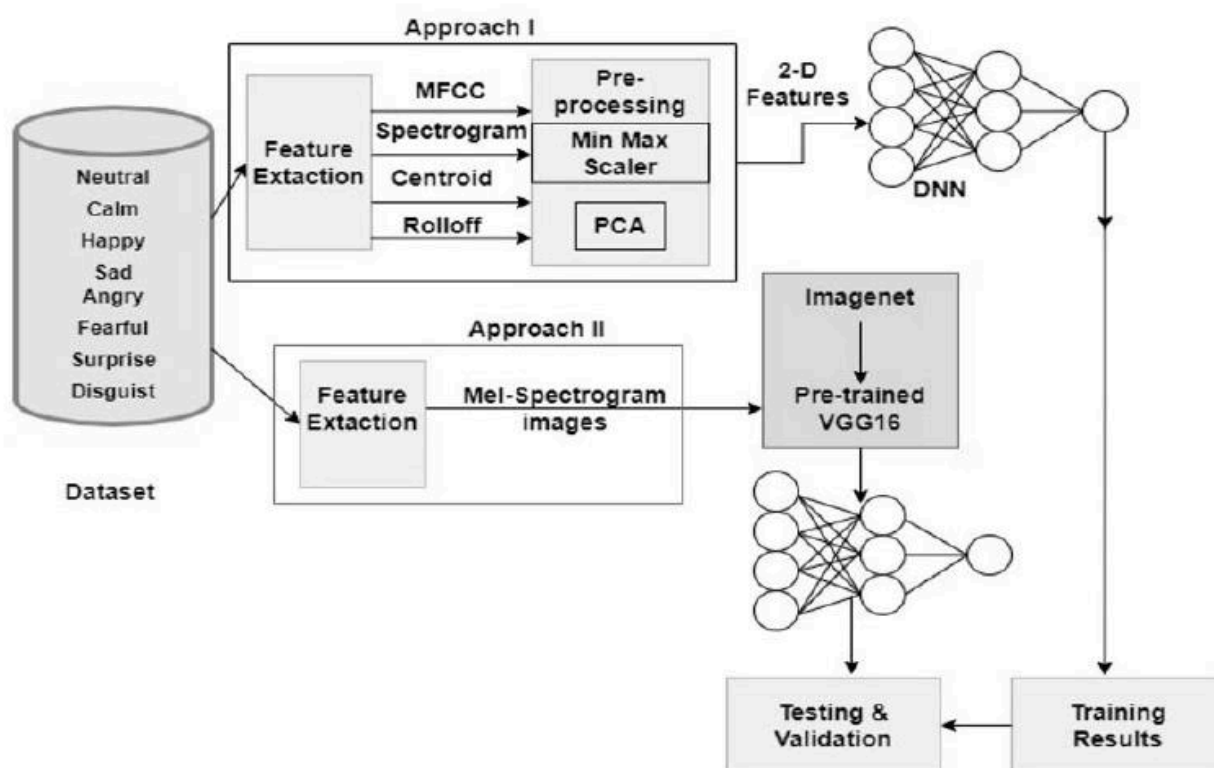


Рис. 1. Схематическое представление двустороннего извлечения признаков из речевых данных [1]

Исследователями работы [2] предложена архитектура, использующая долгую краткосрочную память и полносвязные слои для предсказания эмоций. На вход которой поступают признаки, извлеченные из спектрограммы при помощи сверточной нейронной сети, и вектор, полученный преобразованием мел-кепстральных коэффициентов.

Рассмотренные подходы имеют достаточно высокие значения метрик качества, их существенным недостатком является отсутствие интерпретируемости.

Исследуемые данные. Для обучения модели использовался набор данных Dusha [3] — датасет для распознавания эмоций по речи на русском языке, собранный компанией SberDevices. Данные в нем получены при помощи краудсорсинга: одна группа разметчиков произносила тексты с заданными эмоциями, вторая — ставила метку эмоции на аудиозапись от первой группы, ответы разметчиков агрегировались, в итоге были собраны пары «аудиозапись — эмоция». Множество эмоций состоит из классов: angry, neutral, other, positive, sad. В наборе данных присутствует существенный дисбаланс классов, количество записей каждого класса изображено на рис. 2. Исходил из того, что предсказание каждого класса имеет одинаковую ценность, поэтому в качестве основной метрики использовалась f1-score с макроусреднением.

Извлечение признаков. Перед извлечением признаков к аудио была применена модель улучшения качества звучания голоса — Sepformer [4]. Транскрипция получена путем применения модели распознавания речи whisper-large-v3 [5]. Из транскрипции аудиозаписи были извлечены следующие параметры.

Доля нецензурных слов из всех слов в транскрипции. Гипотеза состоит в том, что использование нецензурных слов тесно связано с сильным выражением эмоций злости и радости, человеку свойственно произносить подобные слова во время гнева или от неожиданной радости.

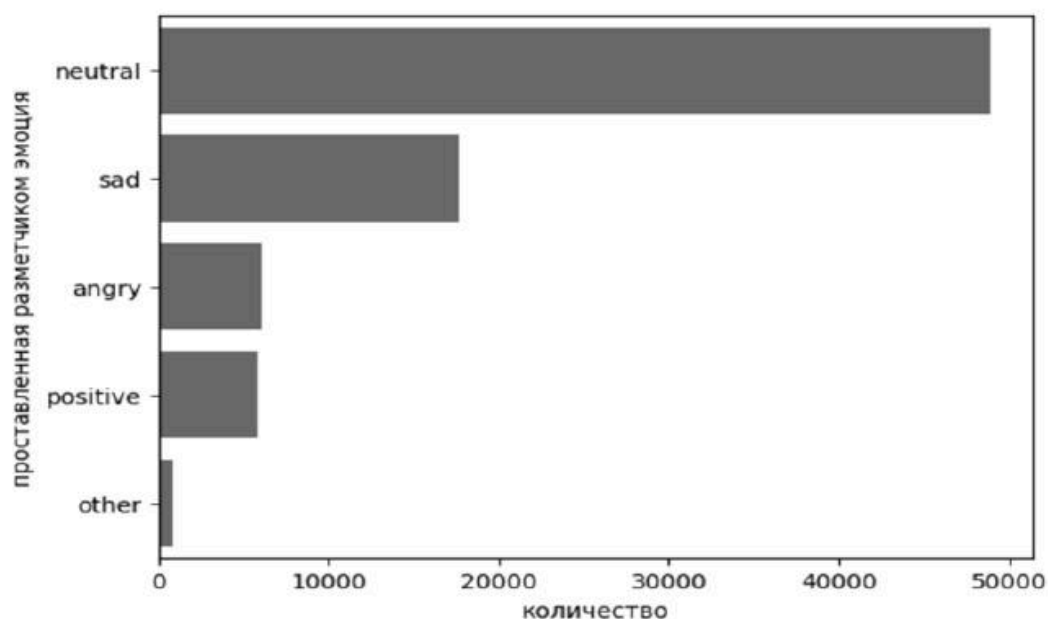


Рис. 2. Количество объектов различных классов

Средняя длина слов. Было сделано предположение, что при испытывании сильной эмоции человек не будет произносить длинные слова.

Средняя скорость произношения букв, слогов, слов. Гипотеза заключается в том, что спикер быстро произносит слова, слоги, буквы, если испытывает эмоции, потому что для него важно скорее донести свою точку зрения.

Средняя значимость слов, рассчитанная при помощи метода TF-IDF. Для каждой части речи подсчитана доля слов, которая к ней относится. Было сделано предположение, что спикеры больше используют глаголы, краткие формы прилагательных и наречий, если испытывают эмоции.

К извлеченным паралингвистическим признакам относятся:

- доля мощности высокочастотных составляющих речи (больше 2000 Гц). Было сделано предположение, что во время радости голос человека становится выше, таким образом, этот параметр поможет предсказывать класс эмоции радости;

- средняя громкость. Гипотеза заключалась в том, что во время сильных эмоций спикер повышает громкость своего голоса, чтобы лучше донести до окружающих свои мысли.

Была подсчитана корреляция Пирсона между признаками. Выявлено, что средняя скорость произношения букв и слов имеют между собой корреляцию больше 0,9, значения корреляции Пирсона с другими признаками для них практически идентичны, что предсказуемо, ведь если человек быстро говорит, то он быстро проговаривает и буквы, и слова. Поэтому из этих скоррелированных признаков оставил один, который имеет наименьшую сумму абсолютных значений корреляции с остальными признаками. Средняя длина слова имеет положительную корреляцию (0,5) со средней скоростью произношения, что говорит о том, что спикеры чаще употребляют длинные слова, когда быстро говорят.

По таблице корреляции Пирсона между извлеченными признаками и классами целевой переменной (рис. 3) видно, что доля ненормативных слов имеет небольшую положительную корреляцию с наличием класса злости, потому что люди часто ругаются во время агрессии. Из таблицы следует, что средняя громкость имеет слабую отрицательную корреляцию с наличием класса other, что говорит о том, что записи с этим классом достаточно тихие, поэтому их сложно отнести к одной из оставшихся эмоций.



Рис. 3. Корреляция Пирсона между извлеченными признаками и классами целевой переменной

Создание модели. На описанных признаках была обучена модель градиентного бустинга при помощи модуля CatBoost для предсказания класса эмоции. Размер тренировочной части — 141 229, тестовой — 13 169. Получившиеся метрики классификации модели можно увидеть в таблице. Видно, что модель значительно лучше справляется с предсказанием эмоции злости и нейтральной эмоции, чем с радостью и грустью.

Классификационные метрики модели градиентного бустинга

	Angry	Neutral	Other	Positive	Sad	Accuracy	Макро-усреднение	Среднее взвешенное
Precision	0,58	0,69	0,76	0,47	0,34	0,68	0,57	0,61
Recall	0,34	0,96	0,49	0,11	0,03	0,68	0,38	0,68
F1-score	0,43	0,80	0,59	0,18	0,05	0,68	0,41	0,60

Модель градиентного бустинга определяет важность весов признаков (рис. 4), из которой следует, что наиболее значимыми признаками являются средняя громкость, средняя значимость слова, доля мощности высокочастотных составляющих речи. Признаки, отвечающие за доли слов, относящиеся к редко встречающимся частям речи, оказались наименее информативными для модели.

Для более подробного анализа влияния признаков на ответ модели была посчитана перестановочная важность. Данный метод для модели показал (рис. 5), что наиболее значимыми являются признаки средней громкости, доля слов с неопределенной частью речи. Оказалось, что важность признаков, полученная данным способом, во многом совпадает с важностью признаков, полученной от самой модели градиентного бустинга.

Значения Шепли в контексте анализа вклада каждого из признаков в предсказание модели — это числа, полученные по формуле (2), отображающие, насколько важен признак для каждого конкретного предсказания. Для этого подсчитывают вклад каждого признака для всех возможных подмножеств без учета рассматриваемого признака из множества признаков:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} (f(S \cup \{i\}) - f(S)), \quad (2)$$

где N — все множество признаков; S — подмножество признаков без i ; $f(S)$ — предсказание модели с признаками S ; $f(S \cup \{i\})$ — предсказание модели с признаками из множества S и i .

Посчитаны значения Шепли для аудиозаписей, где модель верно предсказала класс, построены диаграммы «водопад». На рисунке 6 для аудиозапи-

си с эмоцией гнева изображен вклад признаков в предсказание истинной эмоции, видно, что важнейшие признаки: доля ненормативных слов — в речи действительно слышно нецензурную брань на русском языке, средняя длина слова, доля существительных — они не были употреблены.

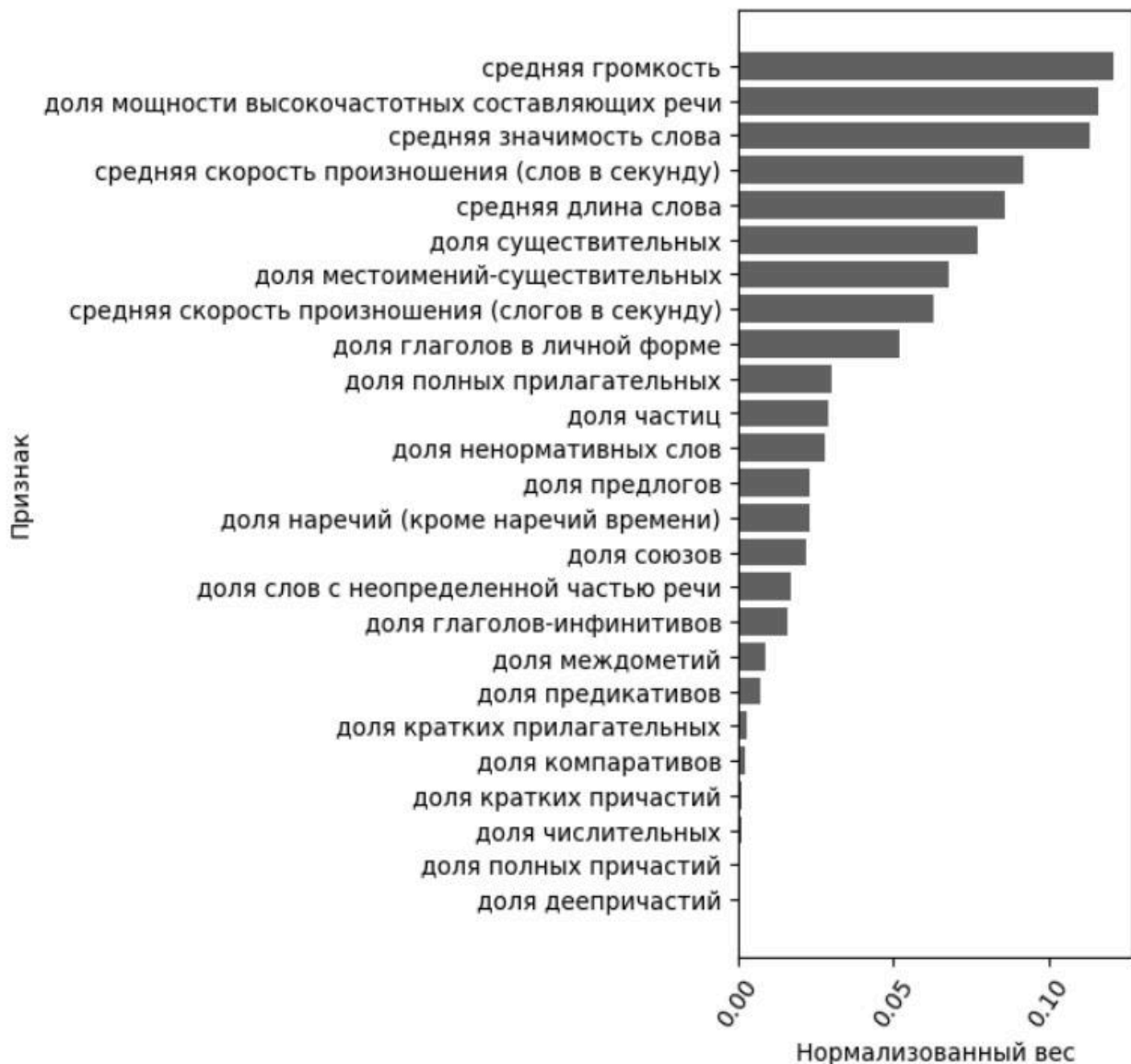


Рис. 4. Важность признаков модели градиентного бустинга

Вклад для каждого конкретного предсказания был усреднен в рамках каждого класса на тестовой выборке, чтобы понять, насколько велик вклад каждого признака для разных классов. На рис. 7 видно, что наиболее важные признаки в среднем для всех классов — средняя громкость, доля высокочастотных составляющих, средняя скорость произношения, средняя значимость слова. Громкость имеет высокое значение для класса позитива, что логично, ведь если человек испытывает радость, он часто повышает голос.

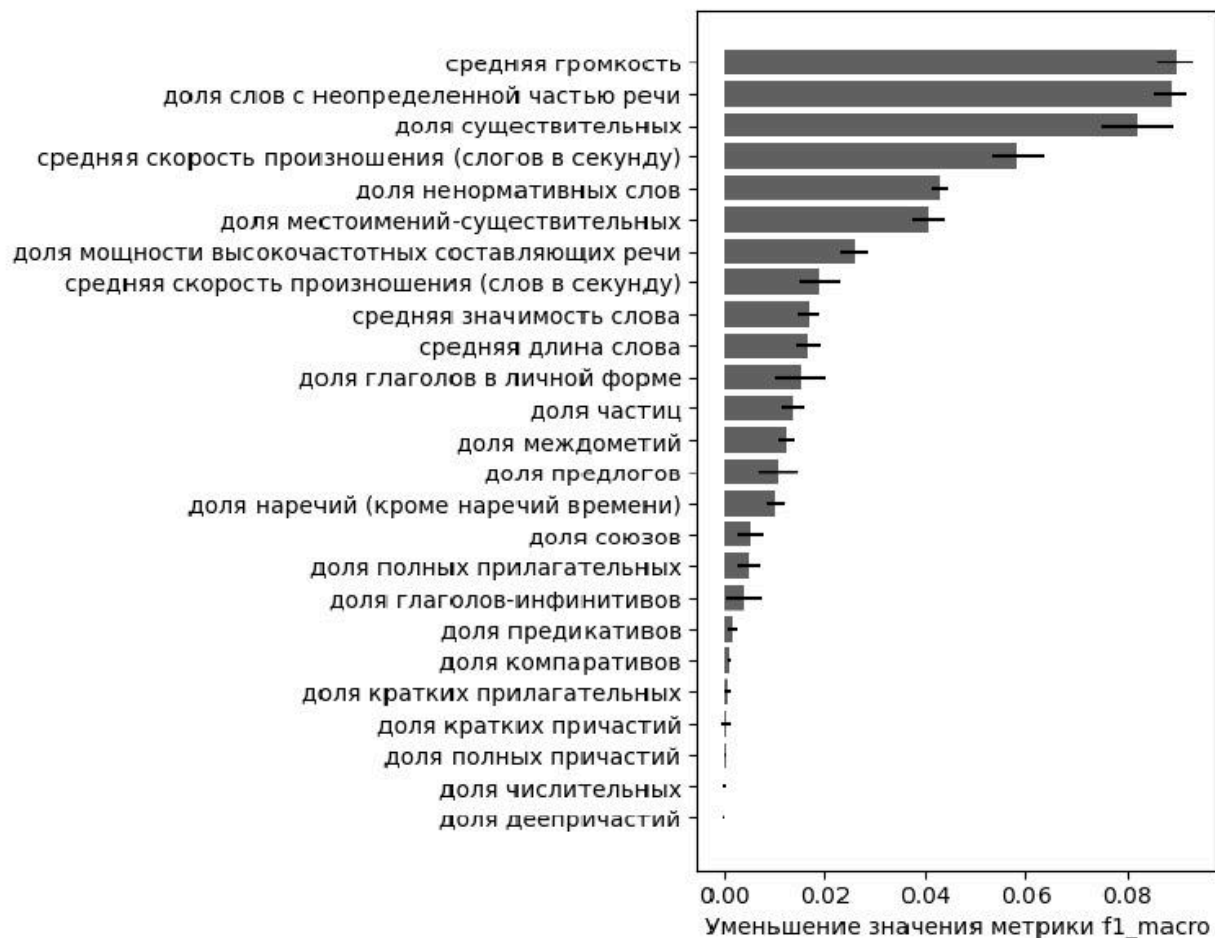


Рис. 5. Перестановочная важность признаков модели градиентного бустинга

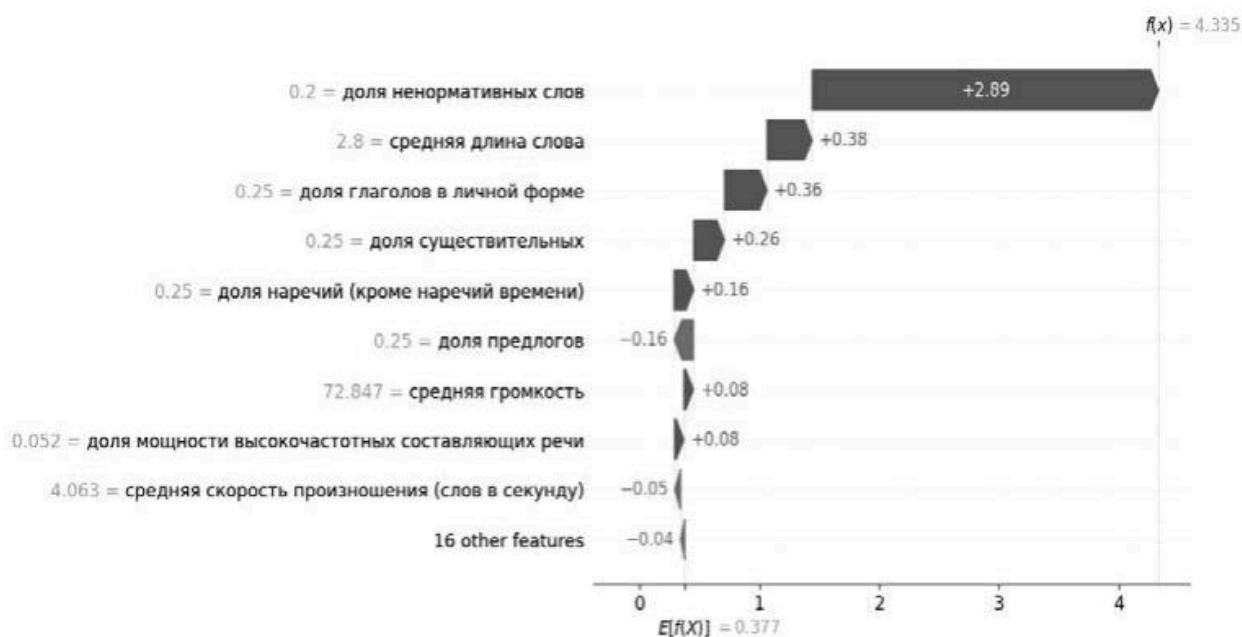


Рис. 6. Вклад каждого признака в предсказание эмоции гнева на отдельной аудиозаписи: $f(x)$ — предсказание модели для выбранного класса; $E[f(x)]$ — базовый (ожидаемый) прогноз, он представляет собой среднее предсказание, которое сделала бы модель, если бы у нее не было информации о конкретном экземпляре

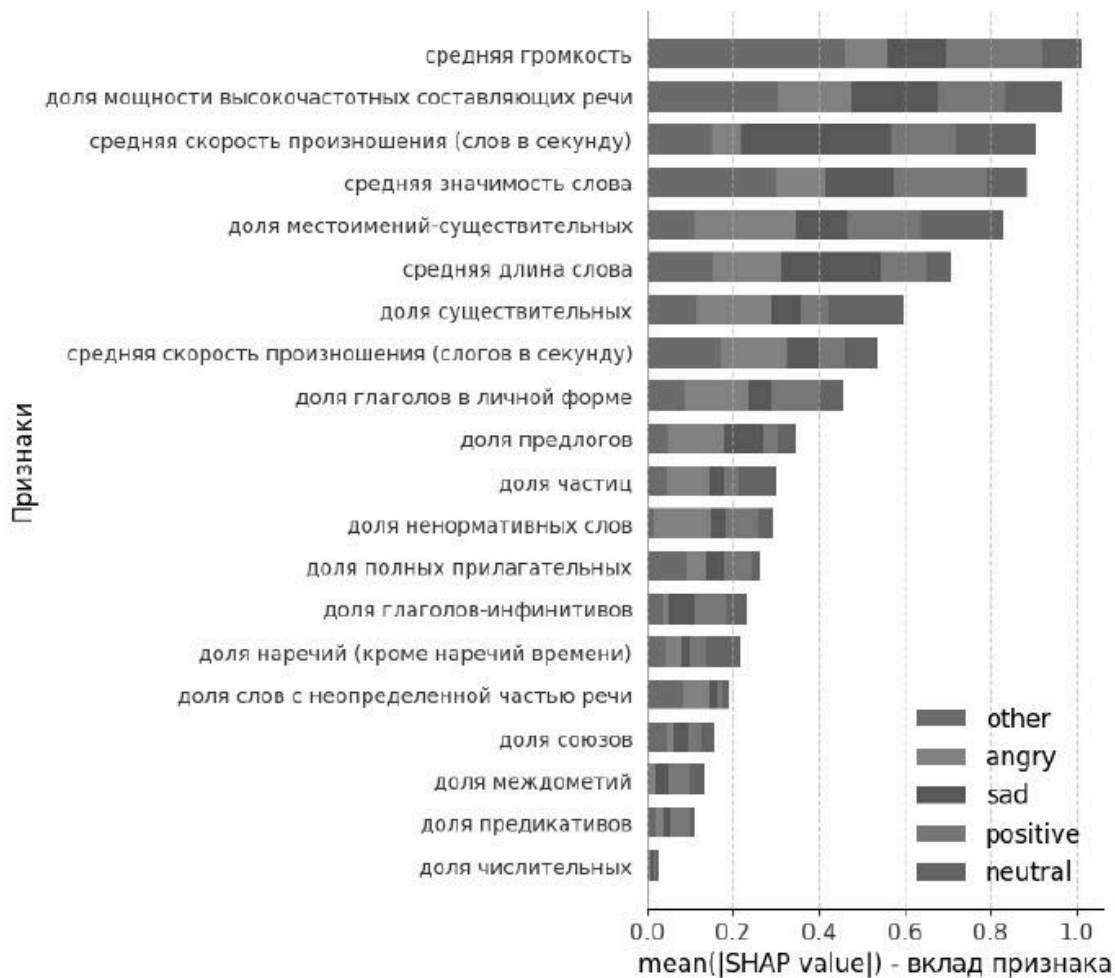


Рис. 7. Важность признаков, основанная на значениях Шепли

Заключение. Лучшая обученная в текущей работе модель имеет метрики accuracy — 0,68, macro f1-score — 0,41, weighted f1-score — 0,6. Лучшая модель на данном датасете на текущий момент — GigaAM, ее метрики: accuracy — 0,90, macro f1-score — 0,84. Заметим, что обученная в работе модель проигрывает модели-нейросети GigaAM по метрикам accuracy и macro f1-score. Однако, в сравнении с GigaAM, главное преимущество обученной модели, что она интерпретируемая.

Выводы о влиянии важнейших признаков на предсказания модели:

- средняя громкость — с ростом значения увеличивается важность для эмоций злости и радости, для нейтральной эмоции зависимость немонотонная;

- доля мощности высокочастотных составляющих речи вносит большой вклад при предсказании всех классов, но для классов грусти и злости его значение чуть больше.

Перспективы дальнейшего исследования:

- изучить, изменяется ли важность признаков для модели, обученной и работающей на других языках;

– использовать метод LIME для более детального описания того, на основе чего модель принимает решение, когда относит аудио к определенному классу.

В дальнейших исследованиях в качестве основы предлагается использовать код текущего исследования для извлечения признаков, построения графиков, обучения моделей, замера их качества.

Литература

- [1] Aggarwal A., Srivastava A., Agarwal A., Chahal N., Singh D., Alnuaim A.A., Alhadlaq A., Lee H.-N. Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. *Sensors*, 2022, vol. 22, iss. 6, art. no. 2378.
- [2] Pulatov I., Oteniyazov R., Makhmudov F., Cho Y.-I. Enhancing Speech Emotion Recognition Using Dual Feature Extraction Encoders. *Sensors*, 2023, vol. 23, art. no. 6640.
- [3] Kondratenko V., Sokolov A., Karpov N., Kutuzov O., Savushkin N., Minkin F. Large Raw Emotional Dataset with Aggregation Mechanism. *ArXiv:2212.12266*, 2022. <https://doi.org/10.48550/arXiv.2212.12266>
- [4] Chen J., Mao Q., Liu D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. *Arxiv:2007.13975*, 2020. <https://doi.org/10.48550/arXiv.2007.13975>
- [5] Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. *Arxiv preprint:2212.04356*, 2022. <https://doi.org/10.48550/arXiv.2212.04356>

The relationship of the emotional coloring of speech with its paralinguistic features and parameters of textual information

Kiselev Viktor Valerievich

Vasilev Aleksandr Viktorovich (*scientific advisor*)

¹ MIPT, Moscow, Russia

² VolSU, Volgograd, Russia

During the work on analyzing emotion recognition data from audio, the relationship of the extracted high-level features to each other and to the classes of the target variable was analyzed, the interpreted model was trained on the most useful parameters, its quality was evaluated, a comparison was made with similar models, and the importance of the features of the best interpreted model was determined.

Keywords: machine learning, emotion recognition, importance of signs, interpretability