УДК 004.934:004.032.26

Литвинов С. В.
магистрант
Московский политехнический университет
Россия, г. Москва
Научный руководитель: Суворов С. В., к.э.н.,
профессор, заведующий кафедрой
Litvinov S. V.
Master's student
Moscow Polytechnic University
Russia, Moscow
Scientific adviser: Suvorov S. V., Candidate of Economic Sciences, Professor, Head of Department

ПОДХОДЫ К РАСПОЗНАВАНИЮ ЭМОЦИЙ В РУССКОЯЗЫЧНОЙ РЕЧИ: АКУСТИЧЕСКИЙ, ТЕКСТОВЫЙ И МУЛЬТИМОДАЛЬНЫЙ МЕТОДЫ

APPROACHES TO EMOTION RECOGNITION IN RUSSIAN SPEECH: ACOUSTIC, TEXTUAL AND MULTIMODAL METHODS

Аннотация: *В статье выполнен обзор и сравнительный анализ подходов к распознаванию эмоций в русскоязычной речи. Рассмотрены три группы методов – акустический, основанный на анализе характеристик речевого сигнала; текстовый, использующий результаты автоматической транскрибации; и мультимодальный, объединяющий оба источника информации. Систематизированы признаки и нейросетевые архитектуры, применяемые в каждом подходе, а также доступные русскоязычные наборы данных. На основе сопоставления по критериям точности, вычислительной стоимости и устойчивости к шуму показано, что наибольший потенциал для русскоязычной речи имеет*

мультимодальный подход, объединяющий акустические и текстовые признаки.

Abstract: *The paper reviews and comparatively analyses the approaches to emotion recognition in Russian speech. Three groups of methods are considered: the acoustic approach, based on the analysis of speech-signal characteristics; the textual approach, which uses the output of automatic speech transcription; and the multimodal approach, which combines both sources of information. The features and neural-network architectures used in each approach are systematised, as well as the available Russian-language datasets. Based on a comparison by accuracy, computational cost and robustness to noise, it is shown that the multimodal approach combining acoustic and textual features has the greatest potential for Russian speech.*

Ключевые слова: *распознавание эмоций по речи, акустические признаки, нейронные сети, мультимодальный анализ, обработка русскоязычной речи, глубокое обучение*

Key words: *speech emotion recognition, acoustic features, neural networks, multimodal analysis, Russian speech processing, deep learning*

Введение

Распознавание эмоционального состояния человека по речи (Speech Emotion Recognition, SER) является одним из активно развивающихся направлений обработки речи и аффективных вычислений. Эмоции проявляются в речи через неязыковые средства – изменение высоты основного тона, интенсивности, темпа и качества голоса, – что позволяет определять состояние говорящего независимо от произносимых слов. Практическая значимость задачи подтверждается её применением в медицинской диагностике, в работе колл-центров, в системах противодействия телефонному мошенничеству и в голосовых ассистентах, адаптирующих ответы под настроение пользователя [3, 4].

Несмотря на значительный прогресс в англоязычном сегменте, для русского языка задача SER остаётся менее изученной: долгое время исследования сдерживались дефицитом крупных открытых размеченных корпусов эмоциональной речи. Появление в 2022 году набора данных Dusha и ряда других ресурсов изменило ситуацию и сделало возможным обучение современных нейросетевых моделей на русскоязычном материале [5]. При этом в литературе отсутствует систематизированное сопоставление применимых подходов именно для русскоязычной речи.

Целью настоящей работы является обзор и сравнительный анализ существующих подходов к распознаванию эмоций в русскоязычной речи – акустического, текстового и мультимодального – с выявлением их сильных и слабых сторон и определением наиболее перспективного направления. Новизна обзора состоит в сопоставлении подходов с учётом специфики русского языка и доступных русскоязычных наборов данных.

Методы и исследования

Материалом обзора послужили научные публикации 2017–2025 годов, посвящённые распознаванию эмоций по речи, а также открытые русскоязычные наборы данных и предобученные модели. Для анализа отбирались работы, в которых приводятся количественные оценки качества классификации эмоций и описание используемых признаков и архитектур. Сопоставление подходов проводилось по следующим критериям: тип используемой информации, применяемые признаки и модели, требования к объёму размеченных данных и вычислительным ресурсам, устойчивость к шуму и применимость к русскоязычной речи.

Задача SER формулируется как многоклассовая классификация: по фрагменту речевого сигнала требуется отнести высказывание к одному из конечного набора эмоциональных классов. Типовой конвейер решения включает сбор и предобработку данных, извлечение признаков, классификацию и постобработку [3]. Основные трудности задачи связаны

с субъективностью эмоций, их распределённостью во времени, межкультурным и индивидуальным разбросом в способах выражения, а также с зависимостью акустических характеристик от пола, возраста и особенностей голоса говорящего [4].

Результаты

Акустический подход опирается исключительно на характеристики речевого сигнала и является исторически первым и наиболее изученным. Выделяют четыре основные группы признаков. Просодические признаки (частота основного тона F0, интенсивность, темп речи, длительность пауз) непосредственно отражают эмоциональную окраску: гневу и страху свойственно повышение F0 и темпа, грусти – их снижение. Признаки качества голоса (джиттер, шиммер, отношение гармоники к шуму) чувствительны к стрессу. Спектральные признаки (спектральный центроид, ширина полосы, спад спектра) и кепстральные коэффициенты (MFCC) компактно описывают тембр и спектральную огибающую сигнала [3].

В рамках классического машинного обучения на векторах этих признаков обучаются метод опорных векторов, случайный лес и скрытые марковские модели. Современные решения используют глубокое обучение: свёрточные нейронные сети обрабатывают мел-спектрограммы как изображения; рекуррентные сети и гибридные архитектуры CNN-BiLSTM моделируют временную динамику; трансформерные модели учитывают глобальный контекст [9]. Отдельный класс составляют самоконтрольные предобученные модели речи – wav2vec 2.0, HuBERT и WavLM, – дообучаемые под задачу SER и обеспечивающие наиболее высокое качество за счёт представлений, полученных на больших массивах непомеченной речи [6, 7, 8]. Для русского языка к ним примыкают модели семейства GigaAM [15]. Преимущество акустического подхода –

независимость от текста и языка; ограничение – чувствительность к шуму и условиям записи.

Текстовый подход предполагает предварительную автоматическую транскрибацию речи в текст с последующей классификацией эмоций по лингвистическому содержанию. На этапе классификации применяются предобученные языковые модели трансформерной архитектуры – прежде всего BERT и его русскоязычные версии (ruBERT), демонстрирующие высокое качество в задачах анализа тональности и классификации текста [10, 11]. Подход хорошо выявляет эмоции, выраженные лексически и семантически, и устойчив к вариативности голоса говорящего.

Вместе с тем текстовый подход имеет принципиальные ограничения. Во-первых, он полностью зависит от качества распознавания речи: ошибки транскрибации, особенно на зашумлённых записях, напрямую снижают точность. Во-вторых, переход к тексту приводит к потере просодической информации – интонации, темпа, тона, – которая зачастую несёт основную эмоциональную нагрузку. Так, одна и та же фраза может быть произнесена с противоположной эмоциональной окраской, неразличимой на уровне текста. Поэтому текстовый подход редко используется изолированно и чаще выступает дополнением к акустическому.

Мультимодальный подход объединяет акустический и текстовый источники информации, компенсируя ограничения каждого из них. Выделяют три стратегии слияния модальностей. При раннем слиянии признаки из обеих модальностей объединяются в единый вектор до классификации. При позднем слиянии независимо обученные модели формируют отдельные предсказания, которые затем агрегируются. Гибридное слияние сочетает оба механизма, в том числе с применением перекрёстного внимания между модальностями [9].

Экспериментальные исследования показывают, что совместный учёт акустических и текстовых признаков систематически повышает точность

по сравнению с одноканальными решениями, поскольку акустика передаёт интонационную составляющую эмоции, а текст – её смысловое содержание. Для русскоязычной речи такая комбинация особенно перспективна: наличие в наборе данных Dusha как аудиозаписей, так и их транскриптов позволяет обучать мультимодальные модели без дополнительной разметки [5]. Основным недостатком подхода является повышенная вычислительная сложность и необходимость согласования двух конвейеров обработки.

Качество моделей SER в первую очередь определяется доступными наборами данных. Для русского языка долгое время единственными ресурсами были RUSLANA (2002) и мультимодальный RAMAS (2018), оба небольшого объёма и с ограниченным доступом [12, 13]. Ситуацию изменил открытый набор Dusha (2022) объёмом около 350 часов записей с транскриптами и разметкой на пять классов, пригодный как для предобучения, так и для дообучения моделей [5]. Для проверки на разговорной речи применяется набор RESD из библиотеки Aniemore [15]. В англоязычном сегменте эталонными остаются RAVDESS, IEMOCAP и CREMA-D, используемые для сравнения [14]. Сводные характеристики наборов приведены в таблице 2.

Таблица 2. Русскоязычные и эталонные наборы данных для задачи SER

Набор данных	Язык	Объём	Классов	Доступ
RUSLANA	русский	≈ 3660 высказываний	6	ограниченный
RAMAS	русский	≈ 7 часов	6	ограниченный
Dusha	русский	≈ 350 часов	5	открытый
RESD (Aniemore)	русский	≈ 3,5 часа	7	открытый
RAVDESS	английский	7356 записей	8	открытый

Сопоставление рассмотренных подходов по ключевым критериям приведено в таблице 1. Акустический подход обеспечивает независимость от языка и текста, но чувствителен к шуму; текстовый устойчив к

вариативности голоса, но теряет просодию и зависит от качества транскрибации; мультимодальный обеспечивает наибольшую точность ценой вычислительной сложности.

Таблица 1. Сравнение подходов к распознаванию эмоций в речи

Критерий	Акустический	Текстовый	Мультимодальный
Источник информации	Речевой сигнал	Транскрипция речи	Сигнал и текст
Типовые модели	CNN, CNN-BiLSTM, wav2vec 2.0, HuBERT	BERT, ruBERT	Гибридные модели со слиянием
Учёт интонации	Полный	Отсутствует	Полный
Устойчивость к шуму	Низкая	Зависит от ASR	Средняя
Вычислительная стоимость	Средняя	Средняя	Высокая
Ожидаемая точность	Высокая	Средняя	Наивысшая

Заключение

Проведённый анализ показывает, что выбор подхода определяется условиями применения. В сценариях с ограниченными вычислительными ресурсами и требованием работы в реальном времени предпочтителен акустический подход на основе компактных свёрточно-рекуррентных моделей. При наличии качественной транскрибации и преобладании лексически выраженных эмоций оправдан текстовый подход. Однако для построения универсального и устойчивого решения для русскоязычной речи наиболее перспективным является мультимодальный подход, что подтверждается как зарубежными исследованиями, так и наличием подготовленных русскоязычных данных с транскриптами.

В работе выполнен обзор и сравнительный анализ трёх групп подходов к распознаванию эмоций в русскоязычной речи – акустического, текстового и мультимодального. Систематизированы применяемые признаки, нейросетевые архитектуры и доступные русскоязычные наборы данных. Показано, что акустический подход остаётся базовым и наиболее

универсальным, текстовый – вспомогательным, а наибольший потенциал для повышения точности имеет мультимодальный подход, объединяющий акустические и текстовые признаки.

Полученные выводы определяют направление дальнейшего исследования – разработку и экспериментальную оценку мультимодальной модели распознавания эмоций для русскоязычной речи на основе набора данных Dusha, а также проверку гипотезы о приросте точности при объединении акустической и текстовой модальностей по сравнению с одноканальными решениями.

Использованные источники:

1. Экман П. Психология эмоций. – СПб.: Питер, 2010. – 334 с.
2. Russell J. A. A circumplex model of affect // Journal of Personality and Social Psychology. – 1980. – Vol. 39, № 6. – P. 1161–1178.
3. Akçay M. B., Oğuz K. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers // Speech Communication. – 2020. – Vol. 116. – P. 56–76.
4. Schuller B. W. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends // Communications of the ACM. – 2018. – Vol. 61, № 5. – P. 90–99.
5. Kondratenko V., Sokolov A., Karpov N. [et al.]. Large raw emotional dataset with aggregation mechanism // arXiv preprint arXiv:2212.12266. – 2022. – 9 p.
6. Baevski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 12449–12460.
7. Hsu W.-N., Bolte B., Tsai Y.-H. [et al.]. HuBERT: self-supervised speech representation learning by masked prediction of hidden units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021. – Vol. 29. – P. 3451–3460.

8. Chen S., Wang C., Chen Z. [et al.]. WavLM: large-scale self-supervised pre-training for full stack speech processing // IEEE Journal of Selected Topics in Signal Processing. – 2022. – Vol. 16, № 6. – P. 1505–1518.
9. Vaswani A., Shazeer N., Parmar N. [et al.]. Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.
10. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
11. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // arXiv preprint arXiv:1905.07213. – 2019. – 7 p.
12. Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing // Speech and Computer (SPECOM). – 2018. – P. 501–510.
13. Makarova V., Petrushin V. A. RUSLANA: a database of Russian emotional utterances // Proceedings of the International Conference on Spoken Language Processing (ICSLP). – 2002. – P. 2041–2044.
14. Livingstone S. R., Russo F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) // PLoS ONE. – 2018. – Vol. 13, № 5. – Article e0196391.
15. Aniemore: открытая библиотека для распознавания эмоций в русской речи : [сайт]. – URL: https://github.com/aniemore/Aniemore (дата обращения: 15.10.2026).