УДК 004.934.2+534.784

АНАЛИЗ ПОДХОДОВ К ПОСТРОЕНИЮ СИСТЕМ РАСПОЗНАВАНИЯ ЭМОЦИЙ ПО РЕЧИ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

Д.В. Краснопрошин
аспирант каф. электронных вычислительных средств, БГУИР
daniil.krasnoproshin@gmail.com

М.И. Вашкевич
д-р техн. наук, доцент, проф. каф. электронных вычислительных средств, БГУИР
vashkevich@bsuir.by

Д.В. Краснопрошин

Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем.

М.И. Вашкевич

Окончил Белорусский государственный университет информатики и радиоэлектроники (БГУИР) в 2008 г., в 2013 г. защитил кандидатскую диссертацию, а в 2022 г. докторскую диссертацию. С 2023 г. работает профессором кафедры электронных вычислительных средств БГУИР. Проводит научные исследования, связанные с применением методов машинного обучения для анализа и классификации речевых сигналов.

Аннотация. В статье представлен обзор методов и алгоритмов распознавания эмоций в речи, охватывающий этапы предобработки данных, извлечения признаков и выбора моделей классификации. Рассмотрены преимущества и недостатки применения статистических моделей. Особое внимание уделено методам с использованием нейронных сетей, анализу их преимуществ и недостатков в контексте распознавания эмоций. Оценены перспективы дальнейших исследований, направленных на улучшение эффективности и интерпретируемости моделей, включая использование мультимодальных данных и сокращение числа обучаемых параметров моделей для повышения производительности.

Ключевые слова: глубокое обучение, распознавание эмоций в речи, сверточные нейронные сети, рекуррентные нейронные сети.

Введение. Распознавание эмоционального состояния по речевому сигналу является важной проблемой, вызывающей все больший интерес у научного сообщества. Это связано с многочисленными прикладными аспектами данного направления, такими как диалоговые системы и помощники, электронное обучение, клинические исследования, распознавание лжи, сфера развлечений, компьютерные игры и колл-центры [1, 2]. Создание эффективных моделей распознавания эмоций в речи может значительно улучшить пользовательский опыт в системах, включающих взаимодействие человека и машины, например, в областях искусственного интеллекта (ИИ) или мобильного здравоохранения (mHealth) [3].

В настоящее время уже предложены различные системы для распознавания в этих областях. Исследователи используют различные методы и классификаторы для определения эмоций, например, метод на основе линейного дискриминантного анализа (ЛДА), метод опорных векторов (МОВ), а также различные архитектуры искусственных нейронных сетей [3].

Тем не менее, данная задача по-прежнему остается весьма сложной, что связано с необходимостью выбора подходящего признакового пространства, наличием фонового шума в аудиозаписях, наличием большого количества естественных языков, а также индивидуальными особенностями говорящих [2].

В данной работе произведен обзор методов обработки речевых сигналов и моделей машинного обучения, которые используются в настоящее время для построения систем распознавания эмоций по речи. Также выполнен обзор наиболее распространенных речевых баз, использующихся для оценки производительности подобных систем. В конце работы сделаны выводы относительно текущего состояния проблемы распознавания эмоций, а также обозначены возможные направления для дальнейших исследований.

Речевые базы данных для задач распознавания эмоций. Исходные наборы данных играют ключевую роль в построении любой системы машинного обучения, в т.ч. систем распознавания эмоций в речи. В настоящее время речевые базы, используемые для обучения, могут содержать как реальные примеры проявления эмоций, так и специально подготовленные записи. Очевидно, что чем более исходные данные ближе к "естественным", тем они сложнее для анализа. Среди наиболее популярных речевых баз данных для распознавания эмоций являются следующие:

1 **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):** Данный набор включает записи от 24 актеров (12 мужчин и 12 женщин), представленных по 104 высказывания на каждого актера (60 речевых и 44 песенных). Набор состоит из 1440 аудиофайлов в wav-формате (16 бит, 48 кГц). RAVDESS содержит в себе различные эмоциональные состояния, такие как нейтральность, спокойствие, счастье, грусть, гнев, страх, удивление и отвращение [4].

2 **Toronto emotional speech set (TESS):** состоит из 2800 аудиозаписей, созданных двумя актрисами в возрасте 26 и 64 лет, которые выражают шесть основных эмоций — гнев, отвращение, страх, счастье, печаль и удивление — вместе с нейтральным эмоциональным состоянием. Каждая запись содержит одно из 200 целевых слов, встроенных в «несущую» фразу «Произнесите слово [цель]», что обеспечивает единообразие в структуре предложения. Аудиофайлы хранятся в wav-формате (16 бит, 24 кГц) [5].

3 **Berlin Emotional Speech Database (Emo-DB):** данный набор включает записи от 10 актеров (5 мужчин и 5 женщин), представленных по 535 высказывания на каждого актера. В базе данных представлены различные эмоции, такие как радость, гнев, грусть, удивление, отвращение, страх, нейтральность и другие [6].

4 **Danish Emotional Speech Database (DES)**: датасет включает записи от 10 актеров (5 мужчин и 5 женщин), представленных по 35 высказываний на каждого актера. В базе данных представлены различные эмоции, такие как радость, гнев, грусть, удивление, страх и нейтральное состояние [7].

5 **Russian Emotional Speech Dialogs with annotated text (RESD):** русскоязычный набор данных эмоциональных речевых диалогов. Этот набор данных был собран из ~3,5 часов живой речи актеров, которые озвучивали заранее распределенные эмоции в диалоге в течение ~3 минут каждый. Эмоции представлены в 7 состояниях: гнев, отвращение, страх, энтузиазм, счастье, нейтральность и грусть [8].

6 **Interactive Emotional Dyadic Motion Capture (IEMOCAP):** мультимодальный набор данных, который состоит из аудио, видео и образцов захвата движения лица, собранных у пяти пар актеров-мужчин и актеров-женщин. Образцы распределены по пяти сессиям, каждая из которых содержит данные от определенной пары. Актеры играли, используя театральные сценарии или импровизировали. Аудиофайлы из данного набора данных разделены на десять классов эмоций: злость, счастье, грусть, нейтральность, разочарование, возбужденность, страх, удивление, отвращение и «другое» [9].

Стоит отметить, что это лишь небольшая часть существующих наборов данных, используемых в данной области. Однако они пользуются наибольшей популярностью и

являются своего рода эталонными при оценке производительности систем распознавания эмоциональных состояний в речи [10].

Распознавание эмоций с использованием классических методов машинного обучения. Для систем распознавания эмоций в речи исходными данными являются аудиозаписи речи. Когда эти записи получены в реальных условиях, необходимо выполнить этап удаления шумов, чтобы улучшить качество данных. В случае с аудиозаписями, полученными в студийных условиях, данный этап можно пропустить. Далее, из обработанных аудиофайлов извлекаются признаки, которые могут быть интерпретированы моделями машинного обучения. Извлечённый набор признаков подается на вход классификатора, в результате чего на выходе модели генерируется метка, соответствующая наиболее вероятной эмоции, определённой на основе анализируемых данных. На рисунке 1 представлен пример описанной выше системы.

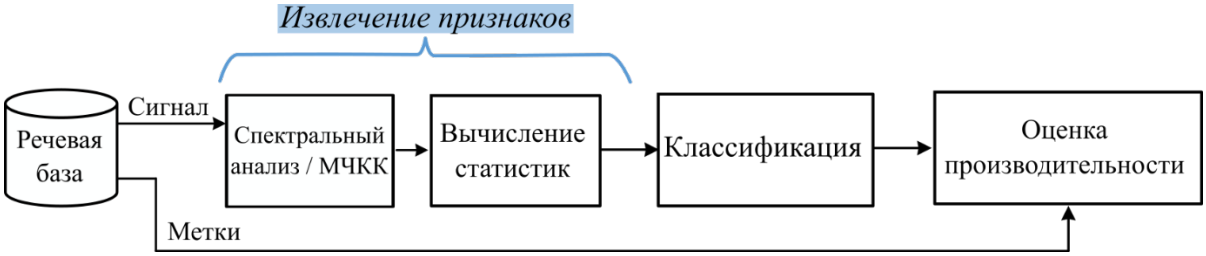

Рисунок 1. Типовая схема системы распознавания эмоций в речи

Большинство современных систем используют мел-частотные кепстральные коэффициенты (МЧКК) в качестве базовых речевых признаков [11]. На основе полученных МЧКК могут также вычисляться первые и вторые производные, а также различные статистики, такие как межквантильный размах, коэффициент асимметрии и эксцесс [Краснопрошин-12]. Вычисление данных характеристик позволяет более точно описать быстроту изменения спектральных характеристик звука во времени. Подобный подход позволяет создать вектор признаков фиксированной длины вне зависимости от длины входного сигнала, что является критическим как для классических моделей машинного обучения (МОВ, ЛДА), так и для различных типов нейронных сетей.

В рамках данного подхода также могут применяться алгоритмы отбора наиболее значимых признаков с целью повысить эффективность классификатора и снизить размерность признакового пространства. Важно подчеркнуть, что, как правило, большинство исследователей используют одни и те же характеристики, но в разном сочетании.

После извлечения и отбора признаков, следующим шагом является выбор подходящего классификатора, который вносит значительный вклад в точность распознавания эмоций. Классические модели такие как МОВ и ЛДА все еще остаются популярными в силу их простоты, неприхотливости к вычислительным ресурсам и хорошей интерпретируемости.

Классические методы классификации речевых эмоций требуют сложной предварительной обработки данных, включая извлечение и отбор признаков. Малые изменения в этих признаках могут значительно повлиять на результаты. Кроме того, увеличение объема данных снижает производительность этих методов.

Применение сверточных нейронных сетей для распознавания эмоций. С развитием методов глубокого обучения, на первый план стали выходить модели на основе различных архитектур нейронных сетей.

Одним из первых типов нейронных сетей, завоевавших популярность среди ученых в области распознавания эмоций в речи, являются различные вариации сверточных нейронных сетей (СНС). Так, например, в работе [13] была предпринята попытка исследовать эффективность СНС в сравнении с иными методами обработки речи, такими

как рекуррентные нейронные сети (РНС) и классические машинные алгоритмы. Из речевого сигнала извлекались пять различных типов признаков такие как МЧКК, мел-спектрограммы, хромаграммы и проч. Затем полученные признаки усреднялись по оси времени и конкатенировались в массив фиксированной длины (*193 признака*), который затем подавался на вход одномерной СНС в качестве входных данных. В работе подчеркивалось, что порядок добавления признаков в итоговый входной вектор имеет критические значение, и что при малейших изменениях результат работы классификатора может поменяться. Выдвигалась гипотеза о том, что смешивание этих признаков во входных данных обеспечивает более разнообразное представление звукового файла, что может привести к лучшему обобщению и лучшим результатам классификации. В качестве финального классификатора был предложен ансамбль моделей на основе СНС. Схематически его работу можно представить следующим образом:

Алгоритм. Ансамбль СНС для классификации эмоции в речевом сигнале.

Вход:

X – исходный вектор признаков;

Результат: метка класса– эмоция, предсказанная классификатором;

Begin:

1: $P(\text{отвращение}) = classifier\_disgust(X)$

2: **if** $(P(\text{отвращение}) \geq threshold)$:

3: return «отвращение»

4: **end if**

5: $P(\text{скука}) = classifier\_boredom(X)$

6: **if** $(P(\text{скука}) \geq threshold)$:

7: return «Скука»

8: **end if**

3: $predicted\_emotion$ = argmax(softmax($multi\_class\_emotion\_classifier(X)$))

4: **return** $predicted\_emotion$

End

Примером еще одной работы по распознанию эмоций с помощью СНС может послужить [14]. В ней предлагается выполнять распознавание на основе кватернионных СНС (*QCNN – Quaternion Convolutional Neural Networks*). В отличие от стандартных вещественных СНС, кватернионные СНС (КСНС) используют многомерные представления данных, что позволяет эффективно кодировать зависимые компоненты акустических признаков.

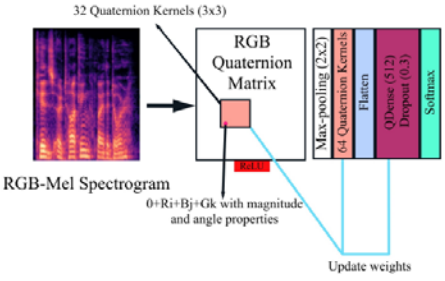

Рисунок 2. Архитектура кватернионной СНС (изображения взято из [14])

В [14] входные данные представлялись в виде многоканальных акустических признаков (например, МЧКК, мел-энергетические спектрограммы), которые обрабатывались в кватернионном пространстве. КСНС позволяют учитывать корреляции между различными акустическими параметрами, что приводит к более информативным представлениям признаков. Архитектура сети включает кватернионные свертки, которые снижают избыточность и улучшают способность модели к обобщению на новых данных (рисунок 2). В [14] экспериментально доказано, что КСНС демонстрирует улучшенную точность распознавания эмоций по сравнению с обычными СНС благодаря более

эффективному представлению взаимосвязанных акустических признаков. При этом удается снизить количество параметров модели, что уменьшает требования к вычислительным ресурсам и снижает вероятность переобучения (оттуда и способность хорошо обобщаться на новые данные). Более того, в ходе экспериментов было установлено, что КСНС могут быть эффективно использованы в условиях ограниченного объема данных, поскольку лучше моделируют зависимости между признаками.

Применение рекуррентных нейронных сетей для распознавания эмоций. Так как речевой сигнал представляет собой последовательность, не менее широкое распространение в задачах распознавания эмоций получили различные модификации РНС. Формально взаимосвязь речевого сигнала и РНС можно описать следующим образом. Речевой сигнал $s(t)$ представляет собой непрерывную функцию времени. В процессе обработки он дискретизируется с частотой f_s, что даёт последовательность дискретных отсчётов:

$$s\lfloor n\rfloor = s(nT_s),\ T_s = \frac{1}{f_s}.$$

В контексте задачи распознавания эмоций на вход РНС обычно подаются временные последовательности акустических признаков, характеризующих речевой сигнал. К таким признакам относятся МЧКК, хроматические признаки, формантные частоты, а также просодические параметры, включая частоту основного тона, интенсивность и прочее. Эти признаки, извлекаемые из фрагментов речевого сигнала, позволяют модели учитывать временные зависимости и динамические изменения в речи, что критически важно для корректной классификации эмоционального состояния говорящего.

Так, с помощью оконного преобразования Фурье и вычисления акустических признаков (например, МЧКК), речевой сигнал преобразуется в последовательность векторов:

$$X = \left[x_{0,}x_1,\ldots,x_T\right],\qquad x_t\ \in \mathbb{R}^d, \tag{1}$$

где сигнал T – длина последовательности, а d – размерность признакового пространства.

РНС моделируют временные зависимости в последовательности X с помощью скрытых состояний h_t, которые передаются между слоями нейронной сети:

$$h_t = f(W_h h_{t-1} + W_x x_t + b_h) \tag{2}$$

где $h_t \in \mathbb{R}^m$ – скрытое состояние на шаге t, $W_h \in \mathbb{R}^{m\times m}$ и $W_x \in \mathbb{R}^{m\times d}$ – матрицы весов, $b_h \in \mathbb{R}^m$ – вектор смещения, $f(\cdot)$ – функция активации.

Сеть обучается так, чтобы скрытые состояния h_t содержали информацию о предыдущих входных значениях, позволяя моделировать контекст в речевом сигнале.

На основе скрытых состояний РНС формируются выходные предсказания, например, вероятность принадлежности речевого фрагмента к определённому эмоциональному классу:

$$y_t = g(V_h h_t + b_y) \tag{3}$$

где V_h – матрица весов, $g(\cdot)$ – функция активации, b_y – вектор смещения.

Схематическое представление РНС показано на рисунке 3.

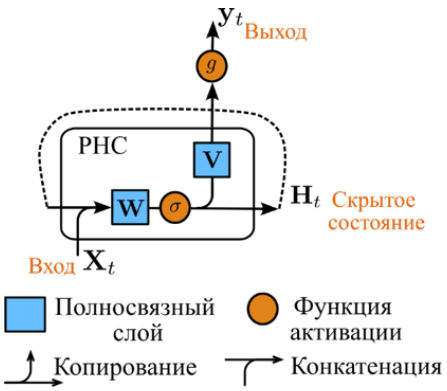

Рисунок 3. Архитектура простой РНС

Базовой моделью для решения задачи классификации эмоций в речи может послужить простая однонаправленная РНС. Выходное состояние такой модели на последнем шаге h_T могут использоваться в качестве входа для функции активации $\mathrm{softmax}(\cdot)$, то есть классификация выполняется по вектору h_T, полученному после обработки всей последовательности входных признаков $X = \{x_0, x_1, \dots, x_T\}$. Формально, вероятность принадлежности к классу c вычисляется как:

$$P(y = c \mid X) = \mathrm{softmax}(V_h h_t + b_y) \quad (5)$$

Таким образом, вся временная информация сжимается в единый вектор h_T, позволяя сети учитывать контекстную информацию. Однако данный подход обладает и рядом ограничений, связанных с неравномерным распределением эмоциональной информации в речевом сигнале. Например, если в начале последовательности присутствует ярко выраженная эмоция (например, гнев или радость), а к концу высказывания речь становится нейтральной, то скрытое состояние h_T будет преимущественно отражать последние, менее эмоциональные фреймы. Это приводит к снижению качества классификации. В дополнение к этому, фреймы речевого сигнала, содержащие тишину или нейтральные интонации, могут затруднять обучение модели, так как финальное скрытое состояние будет неинформативным.

РНС имеет недостаток в виде проблем «исчезающего градиента». Это происходит, когда сеть прекращает обучение в результате изменения выходных данных по отношению к изменению входных данных. В связи с этим начали появляться усовершенствованные модели, такие как сети долгой краткосрочной памяти (*LSTM – Long short-term memory*), а также сети, именуемые управляемым рекуррентным блоком (*GRU – Gated recurrent unit*). Именно эти две усовершенствованные модели стали новым витком в развитии систем распознавания эмоций.

Существует два базовых подхода для построения классификаторов эмоций в речи на основе двунаправленных РНС [15]. Первый и наиболее «наивный» подход заключается в том, чтобы использовать общую эмоцию для каждого фрейма (рисунок 4) и, таким образом, обучать РНС на каждом фрейме и затем с помощью метода обратного распространения ошибки, оптимизируя веса после каждой итерации (в обе стороны). Проблема заключается в том, что не все фреймы в рамках одного высказывания выражают одну и ту же эмоцию. Связано это с тем, что некоторые фреймы могут являться тишиной (молчанием) или просто не содержать никакой эмоции [15]. Более того, поскольку мы предполагаем, что выходные данные РНС являются долгосрочными агрегациями входного сигнала, мы не должны ожидать, что выходные данные будут иметь желаемое долгосрочное представление, начиная с первого фрейма. Наоборот, РНС необходимо получить доступ к истории входного контекста, прежде чем она сможет правильно классифицировать последовательность целиком.

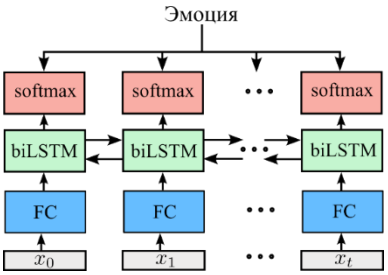

Рисунок 4. Схема с использованием одной метки (эмоции) для каждого фрейма

Альтернативой описанному выше методу является идея использовать только последние скрытые представления (каждого из двух направлений РНС) в качестве входа для функции активации softmax [15] (рисунок 5). Данный способ гарантирует, что модель сможет получить достаточно контекстуальной информации, прежде чем выполнить классификацию. Однако проблема, связанная с тем, что некоторые фреймы могут являться тишиной (молчанием) или просто не содержать никакой эмоции, все еще остается нерешенной. Например, если предложение начинается с ярко выраженной эмоции счастья, но эмоция угасает к концу, выход РНС начнет отклоняться от желаемого представления счастья, поскольку он столкнется с «неэмоциональными» фреймами к концу высказывания. Следовательно, решение полагаться лишь на последнее скрытое состояние РНС не является оптимальным [15].

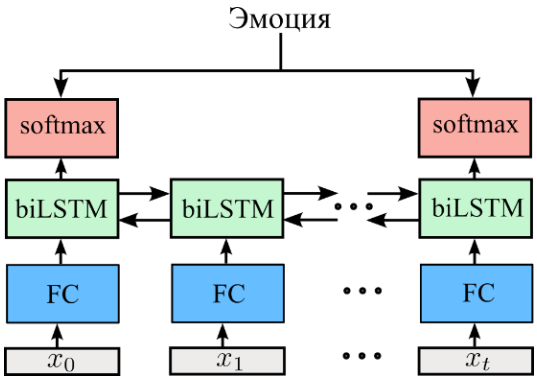

Рисунок 5. Схема с использованием только последних скрытых представлений (каждого из двух направлений РНС) в качестве входа для функции активации

Хорошим примером применения LSTM для задачи распознавания эмоций в речи служит [15], где предлагается новый метод автоматического распознавания эмоций в речи на основе двунаправленной LSTM (*biLSTM – bidirectional LSTM*) с механизмом локального внимания. Механизм внимания позволяет улучшить точность классификации за счет избирательного фокусирования на наиболее информативных участках речевого сигнала, что помогает модели игнорировать менее значимые фрагменты. В качестве входных признаков используются высота тона, вероятность звучания голоса, энергия, скорость пересечения нуля, МЧКК, среднее и дисперсия от МЧКК и др. [15]. Локальный механизм внимания, который позволяет модели фокусироваться только на важных сегментах речевого сигнала, позволяет избежать чрезмерного сглаживания информации, характерного для глобального внимания. Более того, была произведена конкатенация извлеченных признаков и скрытых состояний РНС для улучшенной финальной классификации эмоций. Предложенный подход позволил лучше выделить релевантные сегменты сигнала, а также снизить влияние фонового шума. Тем самым, экспериментально доказано, что локальное (селективное) внимание может применяться для подобных задач. Также в работе [15] отмечено, что локальное внимание способно снижать избыточность обработки, позволяя модели работать более эффективно и интерпретируемо.

Интересный подход был предложен в работе [16], который представляет собой комбинацию СНС и РНС (рисунок 6).

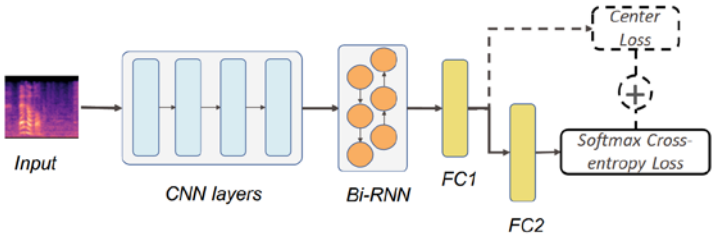

Рисунок 6. Система на основе комбинированного подхода (СНС + РНС) с двумя функциями потерь (изображения взято из [*16*])

В работе решается проблема внутриклассовой изменчивости, возникающей в связи с использованием МЧКК и коэффициентов линейного предсказания и приводящей к снижению эффективности моделей. В качестве альтернативы в статье предлагается использовать мел-спектрограммы, которые позволяют сохранять временные и частотные характеристики речи, в качестве входных данных, которые затем подаются на вход СНС. Наряду с этим вводится дополнительная функция потерь именуемая CenterLoss. Цель данной функции – снизить внутриклассовую изменчивость признаков и тем самым повысить разделимость признакового пространства. В [16] утверждается, что стандартная кросс-энтропийная функция потерь фокусируется на правильной классификации образцов, но не всегда оптимизирует представление данных в признаковом пространстве. Это приводит к высокой внутриклассовой вариативности. Функция CenterLoss решает эту проблему, «притягивая» представления одного класса ближе к его центру в признаковом пространстве. Таким образом, признаки описывающий один и тот же класс становятся более компактными. В итоге, предлагается комбинированный подход, основанный на комбинировании стандартной кросс-энтропийной функция потерь с CenterLoss. Было экспериментально доказано, что данная комбинация усиливает дискриминативную способность модели, что в свою очередь позволяет ей выучить более эффективное признаковое представление входных сигналов.

Итоговая система представляет собой следующее: слои СНС извлекают пространственную информацию из спектрограммы переменной длины, чтобы получить последовательность переменной длины. Двунаправленная РНС (в основе лежит GRU) сжимает последовательность переменной длины до вектора фиксированной длины. Затем первый полносвязный нейронный слой проецирует выходные данные РНС на желаемую размерность. После этого второй полносвязный нейронный слой, выход которого обозначает апостериорные вероятности классов, используется для вычисления softmax энтропийной функции потерь. Эта функция позволяет сети обучаться и выделять разделяющие признаки, а CenterLoss одновременно снижает внутриклассовую вариацию признаков.

В результате экспериментов в [16] делается вывод, что добавление CenterLoss значительно повышает точность классификации за счет уменьшения внутриклассовой изменчивости. Более того, предлагаемый показывает хорошие результаты в условиях ограниченного количества данных, поскольку он помогает извлекать более стабильные признаки. При этом отмечается, что спектрограммы оказываются более информативными, чем МЧКК, так как они сохраняют больше временных и частотных деталей.

Оценка производительности. Оценка производительности моделей классификации эмоций в речи является ключевым этапом в разработке систем автоматического анализа эмоционального состояния и во многом определяет практическую применимость предлагаемых алгоритмов. Для объективного анализа качества работы модели используются различные количественные метрики. Существуют также различные методы разделения данных на тренировочные и тестовые выборки.

Для количественной оценки качества классификации эмоций в речи наиболее часто применяются следующие метрики:

1 **Правильность** (**Accuracy**) – является базовой метрикой, определяющей долю правильно классифицированных примеров относительно общего числа примеров. Однако в задачах с несбалансированными классами данная метрика может быть недостаточно информативной.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \quad (6)$$

где TP – доля истинных положительных результатов (*TP – true positive*), TN – доля истинных отрицательных результатов (*TN – true negative*), FP – доля ложных положительных результатов (*FP – false positive*), FN – доля ложных отрицательных результатов (*FN – false negative*).

Данный параметр применялся в [12, 14] при оценке классификаторов на таких наборах данных как RAVDESS, EMO-DB и IEMOCAP. Значение правильности находится в диапазоне от 0 до 1.

2 **Среднее арифметическое (невзвешенное) полноты (unweighted average recall, UAR).** UAR – это показатель, используемый для измерения общей производительности модели многоклассовой классификации, вычисляет средний уровень запоминания по всем классам, придавая каждому классу одинаковую важность без учета классового дисбаланса:

$$UAR = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{A_{ii}}{\sum_{j=1}^{N_c}A_{ij}} \quad (7)$$

где A – матрица ошибок (confusion matrix); N_c – количество классов.

Значение UAR находится в диапазоне от 0 до 1. Данная метрика использовалась, например, в [12, 17] при оценке классификаторов на основе МОВ и СНС.

Метод разбиения исходного набора данных на тренировочную и тестовую выборки, наряду с выбором метрики качества классификации, может оказать существенное влияние на оценку результатов работы системы распознавания. В связи с этим возникает задача тщательно продумать процесс разделения данных на тренировочные и тестовые выборки. Наиболее распространённым подходом является случайное разбиение данных, например, в пропорции 80% на обучение и 20% на тест, как это было сделано в [12, 15].

Тем не менее, в задачах классификации эмоций в речи необходимо учитывать дикторо-зависимые или дикторо-независимые сценарии. Так в дикторо-независимых сценариях данные одного говорящего не должны присутствовать одновременно в тренировочной и тестовой выборках, чтобы исключить эффект переобучения на индивидуальные особенности голоса. В [12] был предложен подход, призванный устранить данную проблему. В качестве исходного набора данных использовался RAVDESS. Суть подхода заключается в следующем:

1 Набор разбивается на *k* блоков.

2 В цикле для *i* = 1, 2, …, *k* выполняются следующие операции:

– блок *i* устанавливается в качестве тестового набора данных (test data);

– оставшиеся блоки принимаются как тренировочные данные (train data);

– выполняется обучение модели классификатора на тренировочных и оценивается ее эффективность на тестовых данных;

– сохранение результатов классификации для данных из тестового набора;

– сброс параметров модели до исходного состояния для следующей итерации.

3 Расчет оценки эффективности модели на основе сохраненных результатов классификации тестовых данных.

Выбранная стратегия заключается в том, что каждый блок должен содержать одинаковое количество случайно выбранных образцов для каждого класса. При этом должно выполняться условие, что каждый актер представлен либо в обучающей, либо в тестовой выборке, но не в обеих.

Еще одной стандартной техникой для получения более стабильной оценки производительности модели является кросс-валидация (cross-validation) или перекрестная проверка. В задачах классификации эмоций в речи чаще всего применяется k-fold кросс-валидация, при которой данные делятся k равных частей (блоков). На каждом шаге одна из частей используется для тестирования, а оставшиеся — для обучения. Спикер-независимая кросс-валидация требует, чтобы данные одного говорящего полностью присутствовали только в одной из частей, что позволяет оценить обобщающую способность модели на новых говорящих. Например, исходные данные (RAVDESS), согласно схеме, предложенной в [17], разбивали на блоки следующим образом (в скобках указаны номера актеров):

– блок 0: (2, 5, 14, 15, 16);
– блок 1: (3, 6, 7, 13, 18);
– блок 2: (10, 11, 12, 19, 20);
– блок 3: (8, 17, 21, 23, 24);
– блок 4: (1, 4, 9, 22).

Схематическое представление кросс-валидации показано на рисунке 7.

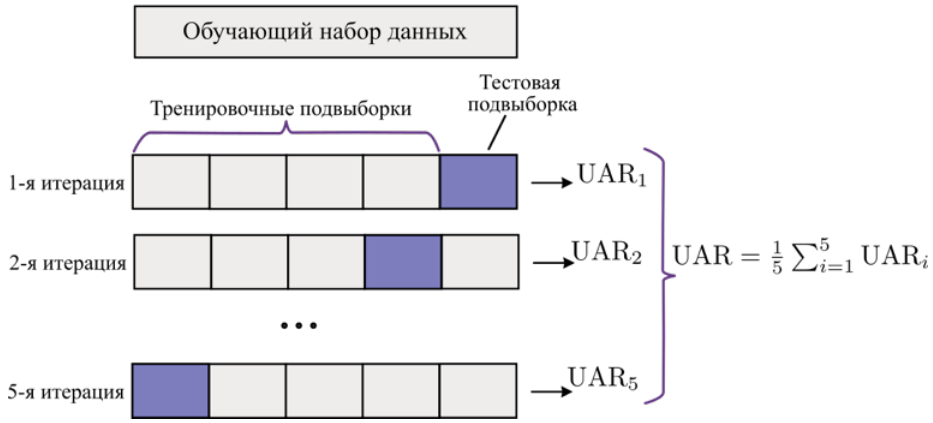

Рисунок 7. Пример кросс-валидации

Заключение. В работе представлен обзор современных систем распознавания эмоций в речи, включая ключевые этапы, такие как предобработка аудиоданных, извлечение признаков и выбор модели классификации. Особое внимание уделено различным архитектурам нейронных сетей, активно применяемым для решения данной задачи. Несмотря на достигнутые успехи, существующие методы, включая традиционные алгоритмы, имеют значительные ограничения, связанные с необходимостью сложной предварительной обработки данных, высокой чувствительностью к изменениям в признаках и снижением производительности при увеличении объема данных. Рассмотренные подходы подчеркивают важность дальнейших исследований для создания более эффективных, интерпретируемых и универсальных моделей. Перспективными направлениями являются разработка устойчивых и масштабируемых методов, таких как нейронные сети, а также интеграция мультимодальных данных для повышения точности распознавания эмоций. Кроме того, важным направлением для будущих исследований является сокращение числа параметров моделей, что позволит улучшить их производительность и снизить вычислительные затраты.

Список литературы

[1] ДИТ Аналитика [Электронный ресурс] / Применение передовых технологий в работе контакт-центра. – URL: https://moscowanalytics.ru/in- dex/research/call_center (дата обращения 12.11.2024).

[2] Han K., Yu D., Tashev I. Speech emotion recognition using deep neural network and extreme learning machine //Proceedings of Interspeech-2014. – 2014. – P. 223-227.

[3] Badshah A. M. et al. Speech emotion recognition from spectrograms with deep convolutional neural network //2017 International conference on platform technology and service (PlatCon). – 2017. – P. 1-5.

[4] Livingstone S. R., Russo F. A. 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. – 2018. – Т. 13. – №. 5. – P. 1-35.

[5] Guilford, J. P., & Schultheiss, H. H. (2008). The Toronto Emotional Speech Set (TESS) [Database]. University of Toronto. – URL: https://utoronto.scholaris.ca/collections/036db644-9790-4ed0-90cc-be1dfb8a4b66. (дата обращения 08.12.2025).

[6] Burkhardt F. et al. A database of German emotional speech // Proceedings of Interspeech-2005. – 2005. – Т. 5. – С. 1517-1520.

[7] Jensen, R., and A. Nielsen. The Danish Emotional Speech Database (DES). Aalborg University, 2010, – URL: https://vbn.aau.dk/en/publications/design-recording-and-verification-of-a-danish-emotional-speech-da. (дата обращения 08.12.2025).

[8] Amentes, Artem, Nikita Davidchuk, and Ilya Lubenets. Russian Emotional Speech Dialogs with Annotated Text (RESD), 2022. – URL https://paperswithcode.com/dataset/resd. (дата обращения 08.12.2025).

[9] Busso C. et al. IEMOCAP: Interactive emotional dyadic motion capture database // Language resources and evaluation. – 2008. – Vol. 42. – P. 335-359.

[10] Видман В. В. и др. Распознавание эмоций из речевого сигнала // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых учёных, 3-7 декабря 2018 г., г. Томск.—Томск, 2019. – 2019.

[11] Huang X. et al. Spoken Language Processing: A guide to theory, algorithm, and system development. – Prentice hall PTR, 2001.

[12] Краснопрошин Д. В., Вашкевич М. И. Метод распознавания эмоций в речевом сигнале с использованием машины опорных векторов и надсегментных акустических признаков //Доклады БГУИР. – 2024. – Т. 22. – №. 3. – С. 93-100.

[13] Issa D., Demirci M. F., Yazici A. Speech emotion recognition with deep convolutional neural networks //Biomedical Signal Processing and Control. – 2020. – Т. 59. – С. 101894.

[14] Muppidi A., Radfar M. Speech emotion recognition using quaternion convolutional neural networks // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2021. – P. 6309-6313.

[15] Mirsamadi S., Barsoum E., Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention // Proceedings of IEEE International conference on acoustics, speech and signal processing (ICASSP). – 2017. – P. 2227-2231.

[16] Dai D. et al. Learning discriminative features from spectrograms using center loss for speech emotion recognition // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). –2019. – P. 7405-7409.

[17] Luna-Jiménez C. et al. Multimodal emotion recognition on RAVDESS dataset using transfer learning // Sensors. – 2021. – Т. 21. – №. 22. – P. 1-29.

Авторский вклад

Краснопрошин Даниил Вадимович – решение задач исследования, анализ полученных результатов, формирование структуры статьи.

Вашкевич Максим Иосифович – постановка цели и задач исследования, руководство исследованием, анализ полученных результатов, редактирование статьи.

ANALYSIS OF APPROACHES TO BUILDING SPEECH EMOTION RECOGNITION SYSTEMS USING DEEP LEARNING METHODS

D.V. Krasnoproshin
PhD Student at the Department of Electronic Computing Facilities, Belarusian State University of Informatics and Radioelectronics

M.I. Vashkevich
Dr.Sc., Professor at the Department of Electronic Computing Facilities, Belarusian State University of Informatics and Radioelectronics

Abstract. The article provides an overview of methods and algorithms for speech emotion recognition problems, covering the stages of data preprocessing, feature extraction, and classification model selection. The application of statistical models and their limitations are discussed. Special attention is given to neural network technologies, analyzing their advantages and disadvantages in the context of solving the problem. The prospects for further research aimed at improving the efficiency and interpretability of models are evaluated, including the use of multimodal data and dimensionality reduction to enhance performance.

Keywords: deep learning, speech emotion recognition, convolutional neural networks, recurrent neural networks.