Litera
Правильная ссылка на статью:
Лемаев В.И., Лукашевич Н.В. Автоматическая классификация эмоций в речи: методы и данные // Litera. 2024. № 4. DOI: 10.25136/2409-8698.2024.4.70472 EDN: WOBSMN URL: https://nbpublish.com/library_read_article.php?id=70472

Автоматическая классификация эмоций в речи: методы и данные

Лемаев Владислав Игоревич

аспирант, кафедра теоретической и прикладной лингвистики, Московский Государственный Университет

119991, Россия, г. Москва, Ленинские Горы, 1с51

vladzhkv98@mail.ru

Лукашевич Наталья Валентиновна

доктор технических наук

профессор, кафедра теоретической и прикладной лингвистики, Московский Государственный Университет имени М.В. Ломоносова

119991, Россия, г. Москва, Ленинские Горы, 1с51, ауд. 953

louk_nat@mai.ru

Статья из рубрики "Автоматическая обработка языка"

DOI:

10.25136/2409-8698.2024.4.70472

EDN:

WOBSMN

Дата направления статьи в редакцию:

14-04-2024

Дата публикации:

21-04-2024

Аннотация: Предметом настоящего исследования являются данные и методы, применяемые в задаче автоматического распознавания эмоций в разговорной речи. Данная задача приобрела в последнее время большую популярность, в первую очередь благодаря появлению больших датасетов размеченных данных и развитию моделей

машинного обучения. Классификация речевых высказываний обычно осуществляется на основе 6 архетипических эмоций: гнева, страха, удивления, радости, отвращения и грусти. Большинство современных методов классификации основано на машинном обучении и модели трансформера с использованием подхода самообучения, в частности, такие модели, как Wav2vec 2.0, HuBERT и WavLM, которые рассмотрены в данной работе. В качестве данных анализируются размеченные английские и русские датасеты эмоциональной речи, в частности, датасеты Dusha и RESD. В качестве метода был проведён эксперимент в виде сравнения работы моделей Wav2vec 2.0, HuBERT и WavLM на относительно недавно собранных русских датасетах эмоциональной речи Dusha и RESD. Основной целью работы выступает анализ доступности и применимости имеющихся данных и подходов распознавания эмоций в речи для русского языка, исследований для которого до этого момента было проведено сравнительно мало. В рамках проведённого эксперимента были получены хорошие результаты качества классификации эмоции на русских датасетах Dusha и RESD. Наилучший результат продемонстрировала модель WavLM на датасете Dusha - 0.8782 по метрике Accuracy. На датасете RESD лучший результат тоже получила модель WavLM, при этом для неё было проведено предварительное обучение на датасете Dusha - 0.81 по метрике Accuracy. Высокие результаты классификации, в первую очередь за счёт качества и объёма собранного датасета Dusha, готворят о перспективности дальнейшего развития данной области для русского языка.

Ключевые слова:

обработка естественного языка, распознавание эмоций, распознавание речи, машинное обучение, трансформеры, Wav2vec, HuBERT, WavLM, Dusha, RESD

Введение

Целью создания систем распознавания эмоций является определение внутреннего состояния говорящего посредством анализа его речи. Такие системы широко применяются как в областях взаимодействия человека с компьютером, так и в различных системах безопасности (например, при установлении эмоционального состояния водителя машины).

Задача автоматического распознавания эмоций содержит множество трудностей, которые для не знакомого с областью человека могут оказаться неочевидными. Так, речь и выражение эмоций варьируются в зависимости от многих параметров, будь то разные возрастные группы говорящих, разные языки, разный культурный бэкграунд. При этом не очень понятно, какие из акустических признаков в таком случае будут наиболее важны для классификации эмоции. Более того, в одном речевом высказывании может содержаться сразу несколько выраженных эмоций, каждая из которых относится к отдельному речевому фрагменту.

Создание системы автоматического распознавания эмоций делится на два основных этапа – сбор обучающих данных и создание алгоритма классификации эмоций. Первый этап на текущий момент относительно решён для английского языка, для которого существует множество работ по этой теме и свободно доступно значительное количество обучающих данных, но для других языков по-прежнему наблюдается недостаток размеченных данных. Подходы к решению второго этапа, однако, не прекращают активно развиваться и по сей день. Наиболее перспективными на данный момент

являются появившиеся недавно модели на основе архитектуры трансформера, такие как Wav2Vec [1], которые позволяют автоматически анализировать имеющиеся данные и на их основе уже самостоятельно классифицировать новые, ранее не встречавшиеся высказывания.

В статье описывается задача автоматического распознавания эмоций и подходы к её решению. Цель статьи – сравнить между собой результаты часто применяемых методов на основе транформеров на материале русского языка. В разделе 1 дано описание инвентарей эмоций, на основе которых производится классификация. Раздел 2 посвящен сбору и обработке данных, а также имеющимся на данный момент корпусам эмоциональной речи. В разделе 3 приведены методы классификации, которые активно применяются исследователями в своих работах, в том числе и в этой статье. Наконец, в разделе 4 описано тестирование и сравнение результатов классификации эмоций, полученных разными моделями.

1. Составление инвентаря эмоций

Одной из первостепенных проблем области является определение инвентаря эмоций, на основе которого должна проводиться классификация. Само понятие эмоции до сих пор не имеет согласованного теоретического определения, а те эмоции, которые выражены в речи, имеют высокие вариативность и наложение. Типичные наборы эмоций, используемые в теоретических и практических областях психологии, насчитывают около 300 разнообразных эмоций и их вариаций. Однако это число чрезвычайно сложно реализовать в задаче автоматической классификации. Поэтому исследователи обычно используют набор из 6 *основных* эмоций, на которые часто принято раскладывать все остальные эмоции — это гнев, страх, удивление, радость, отвращение и грусть [2]. Эти эмоции являются универсальными для большинства языков и носят название *архетипических*. Большинство датасетов используют данный набор эмоций, часто так или иначе его модифицируя [3, 4].

Помимо классификации на основе набора эмоций, исследователи также применяют подход классификации на основе двумерного пространства. Первой осью такого пространства выступает понятие «валентности», которое описывает выражаемую эмоцию в пространстве «положительность-отрицательность». Второй осью выступает понятие «возбуждения», которое описывает степень проявления выражаемой эмоции. Такое двумерное пространство в высокой степени коррелирует с названными выше архетипическими эмоциями. Так, возбуждение симпатической нервной системы происходит при выражении эмоций гнева, страха и радости и приводит к повышенному сердечному ритму, напряжению мышц, повышенному кровяному давлению и сухости во рту, быстрой и энергичной речи. Напротив, выражение эмоции грусти приводит к возбуждению парасимпатической системы и проявляется в пониженном сердечном ритме и кровяном давлении, что в свою очередь отражается в медленной и тихой, часто низко-тональной речи. Эти речевые характеристики и используются для оценки степени проявления эмоции по оси «возбуждения». Однако, хоть степень активации энергии и помогает разграничивать эмоции грусти и радости, она не даёт возможности разграничивать эмоции радости и гнева. Для этой цели и используется пространство «валентности», но, к сожалению, у исследователей нет чёткого понимания того, с какими из акустических характеристик оно коррелирует. Последние исследования показывают, что применение трансформеров позволяет улучшить качество предсказания по шкале «валетности» за счёт имплицитной лингвистической информации [5].

2. Данные для распознавания эмоций в речи

Одной из самых важных задач при проектировании подобных систем является сбор обучающих данных, большой объём которых необходим для методов на основе машинного обучения, активно применяющихся в данный момент. Базы данных эмоциональных высказываний можно категоризовать в терминах естественности, эмоций, говорящих, языка и т. д. Естественность является одним из важнейших параметров, который следует учитывать при сборе высказываний и создании базы данных. Можно выделить 3 степени естественности речи: естественная, полуестественная и искусственная. В базах данных с естественной речью [6] материал обычно собирается из реальных бытовых ситуаций с целью записи реальных эмоций. Подобные данные, однако, довольно редко используются в реальных исследованиях вследствие законных и моральных ограничений, которые могут возникнуть при таком подходе. Поэтому для исследований речевой материал обычно специально записывается отдельно с привлечением добровольцев.

В базах данных с полуестественной речью применяется 2 подхода: на основе сценария (например, [4]) и на основе актёрской игры [7]. При первом подходе говорящий сначала должен войти в нужное эмоциональное состояние (радость, гнев и т.д) – это может быть выполнено посредством переживания различных воспоминаний, компьютерных игр или диалога с компьютером (последний метод носит название «Волшебник страны Оз»). Затем говорящему дают для прочтения заранее написанный текст, который по содержанию соответствует его эмоциональному состоянию. При втором же подходе языковые данные извлекаются из фильмов и радио-постановок.

Наконец, в базах данных с искусственной речью используются заранее написанные тексты (обычно в нейтральном стиле), которые даются на прочтение с заданной эмоцией профессиональным актёрам. Основной проблемой при таком подходе часто является наигранность различных эмоций. Для её решения записи порой проводят с привлечением людей, не обладающими профессиональными актёрскими навыками. Например, в базе данных датской эмоциональной речи [8] были записаны высказывания полупрофессиональных актёров, чтобы эмоции были более приближенными к реальным речевым ситуациям.

Помимо актёрского опыта важно также учитывать другие характеристики говорящих, а именно возраст, пол, родной язык. Мужчины и женщины, а также старики и подростки по-разному выражают в своей речи эмоции, поэтому для большинства датасетов, целью которых стоит универсальность и применимость для различных исследований, важно сбалансировать говорящих по полу и возрастным группам, чтобы классифицирующий алгоритм мог распознавать их с одинаковой уверенностью.

Родной язык говорящих обычно соответствует тому языку, для которого и создаётся датасет. Однако даже в пределах одного языка существует множество диалектов и локальных говоров, но их обычно не включают в датасеты, за исключением направленного исследования. Помимо этого, также создаются мультиязычные датасеты с целью изучения выражения эмоций в разных языках и создания универсальной модели [9].

При создании корпуса также важно учитывать длительность высказываний и их количество. Так, в общем случае высказывания не должны быть слишком длинными или слишком короткими. Средняя длина высказывания в идеале должна составлять около 7

секунд (в пределах от 3 до 11 секунд [10]). При слишком большой длине эмоции могут чередоваться или сменяться — особенно при извлечении высказываний из фильмов или телепередач. При слишком маленькой же длине в высказывании будет слишком мало признаков, которые необходимы для классификации выражаемых эмоций. Однако, хоть такие данные и помогают добиться хороших результатов распознавания эмоций в контролируемой среде, они плохо отражают реальную картину их выражения. Для повышения валидности, датасет может быть составлен из несбалансированных по длине высказываний, но для хороших результатов требования по объёму к нему будут выше.

Также важно учитывать и распределение высказываний по классам. Для должной оценки качества распознавания эмоций, количество высказываний для каждой из классифицируемых эмоций должно быть примерно одинаковым. Однако в реальности частота встречаемости каждой эмоции зависит от условий, в которых происходит общение. Так, при извлечении данных из телепередач, особенно тех, что касаются социальных или политических тем, превалирующим будет количество высказываний с негативными эмоциями [11]. При записи же диалога в контролируемых условиях между мало знакомыми участниками – которые, например, пытаются совместными усилиями решить данную им задачу, – высказывания будут носить преимущественно положительный или нейтральный эмоциональный окрас.

Наконец, важно учесть лексическое разнообразие высказываний, то есть количество в них уникальных слов. Это особенно необходимо при подходе с прочтением предложений с заранее заданными эмоциями. Часто в таких случаях одни и те же по лексическому составу предложения озвучиваются несколько раз с разными эмоциями, поэтому лексическое разнообразие может быть низким, что снижает дальнейшую результативность классификатора. Важно составить несколько наборов предложений, которые были бы как можно более лексически разнообразны.

Для английского языка наиболее популярными являются датасеты IEMOCAP и RAVDESS. Для русского языка среди имеющихся на данный момент в открытом доступе датасетов следует, в первую очередь, выделить датасеты Dusha и RESD. Рассмотрим далее доступные датасеты подробнее. Таблица 1 содержит характеристики наиболее известных датасетов.

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database)[4] – это мультимодальный датасет, созданный для исследования коммуникации между людьми и состоящий из более 12 часов аудио и видеозаписей разговоров между актерами. Датасет включает в себя 5 диалогов между актёрами (по 2 актёра в каждом разговоре), в записях участвовали 5 мужчин и 5 женщин. Актёров попросили сыграть заранее составленные реплики, а также сымпровизировать диалог в предлагаемом сценарии, с целью вызвать определённые эмоции. Каждая сессия записывалась с помощью специальной системы захвата движений, что позволило получить не только аудиозапись, но и информацию о движениях лица, головы и рук. Всего представлено 5 классов эмоций: Радость, Гнев, Грусть, Отвращение и Нейтральное состояние. Данный датасет в настоящий момент является наиболее популярным в задачах автоматического распознавания эмоций.

RAVDESS [12] – один из популярных датасетов, используемый в задачах распознавания эмоций в речи и пении. Датасет имеет большой объём (7356 высказываний) и содержит аудио и видеозаписи профессиональных актеров, проживающих в Северной Америке. Участники зачитывали и пели отведённые им предложения. В общей сложности в

датасете присутствуют записи 24 актеров (12 мужчин и 12 женщин). Для речевых записей выделялись эмоции Спокойствия, Радости, Грусти, Гнева, Страха и Отвращения. Для записей с пением – Спокойствия, Радости, Грусти, Гнева и Страха. Отличительной чертой данного датасета является то, что каждая эмоция имеет две степени интенсивности – слабую и сильную. RAVDESS широко используется в исследованиях и разработке систем распознавания эмоций. Он предоставляет хороший набор данных для тренировки и тестирования моделей машинного обучения, которые занимаются анализом эмоций на основе аудио и видеоданных.

Датасет **Dusha** [3] создан компанией SberDevices и является на данный момент крупнейшим датасетом на русском языке, предназначенным для решения задач распознавания эмоций в разговорной речи. Датасет разделён на 2 части: для первой части под названием Crowd авторы сгенерировали тексты на основе общения реальных людей с виртуальным ассистентом, которые затем были озвучены с помощью краудсорсинга (говорящим давались текст и эмоция, с которой данный текст должен был быть произнесён, и затем полученные аудиозаписи дополнительно проверяла вторая группа); вторая часть под названием Podcast содержит короткие (до 5 слов) отрывки из русскоязычных подкастов, которые были затем классифицированы по эмоциям. Всего представлено 5 классов эмоций: Гнев, Грусть, Позитив, Нейтральная и Другое. Полученный в итоге датасет содержит около 300 тысяч аудиозаписей длительностью около 350 часов и их письменные расшифровки.

RESD (Russian Emotional Speech Dialogs)[13] – русскоязычный датасет, представленный в открытой библиотеке Aniemore, предназначенной для обработки и распознавания эмоциональной окраски разговорной и письменной речи. Для записи были составлены диалоги с заранее заданными эмоциями, которые затем были озвучены актёрами. Всего датасет содержит около 4 часов записанной речи (около 1400 аудиозаписей) и соответствующую письменную расшифровку каждой аудиозаписи. Классификация произведена на 7 эмоций: Нейтральная, Гнев, Энтузиазм, Страх, Грусть, Радость и Отвращение.

Название датасета	**Объём**	**Количество говорящих**	**Метод создания**	**Типы данных**	**Языки**	**Класс на э (вк нейтр сост**
IEMOCAP	151 запись (более 12 часов)	10	Прочтение реплик и импровизация (диалог)	Аудио и видео	Английский	Гнев, Гр Отвр
RAVDESS	7356 высказываний	24	Прочтение реплик	Аудио и видео	Английский (Северо-американский акцент)	Гнев, Гр Отвр С Уди Спок
Dusha	300000 записей (350 часов)	Более 2000	Прочтение реплик / извлечение из подкастов	Аудио	Русский	Гнев Позити
						Гнев,

RESD	1400 записей (4 часа)	Около 50	Прочтение диалогов	Аудио	Русский	Страх Ра Отвр

Таблица 1. Характеристики датасетов

3. Методы классификации

Ранние подходы к задаче автоматического распознавания эмоций представляли собой в основном ручное извлечение признаков, которые затем подавались на вход классическим алгоритмам классификации, таким как SVM [14] и Random Forest [15]. Наибольшей популярностью пользовались спектральные признаки. Одними из них были MFCC-признаки [16], которые используются и в современных моделях, например, в HuBERT (см. ниже). MFCC-признаки моделируют характеристики голосового тракта говорящего, тем самым позволяя анализировать произносимый в конкретный момент времени звук.

Следующим важным этапом стало развитие нейронных сетей в области обработки естественного языка, которые сначала продемонстрировали хорошие результаты в задачах классификации текстов, а затем привлекли к себе внимание и в задаче автоматического распознавания эмоций [17]. Подход с использованием модели LSTM [18] имел на тот момент наибольшую популярность.

Последние же годы на первый план вышел подход на основе самообучения (Self-Supervised Learning; SSL). Он позволяет алгоритмам самостоятельно находить закономерности в исходных данных и тем самым позволяет достичь более высоких результатов. При таком подходе открываются возможности использования огромных объёмов данных для обучения, вплоть до всех данных, доступных в интернете. Это стало возможно в первую очередь благодаря появлению архитектуры трансформера [19], которая, в отличие от предшествующих моделей нейронных сетей, позволила параллелизировать обработку данных при обучении. Большинство моделей, созданных на основе самообучения, лежат в открытом доступе и для хороших результатов перед применением требуют лишь небольшого дополнительного дообучения на данных из соответствующей области.

Одной из наиболее известных моделей самообучения, созданной на основе архитектуры трансформера, является BERT [20], который был создан исследователями корпорации Google с целью обработки письменного языка и активно применяется при формировании ответов на запросы в одноимённом поисковике. Однако, для обработки звучащей речи BERT подходит не так хорошо, как для обработки письменной. Во-первых, BERT требует на вход отдельные токены, то есть слова или их части. Для речи в таком случае необходимо использовать токенизатор, а это дополнительная задача, при выполнении которой в полученных данных будет много шума. Во-вторых, сам BERT обучен на большом корпусе текстовых данных, и их стиль плохо подходит для разговорной речи.

Для задач обработки разговорной речи были предложены специальные модели. Среди них наиболее популярными на данный момент являются Wav2vec 2.0 [21], HuBERT [22] и WavLM [23]. Все модели работают непосредственно с необработанными исходными аудиозаписями, разбивая их на “фреймы” – короткие отрезки, длительностью обычно 20 мс.

Wav2vec 2.0 случайным образом маскирует некоторые фреймы аудиозаписи, поступающей на вход, и учится предсказывать эти замаскированные фреймы, тем самым получая качественную репрезентацию аудиоданных без дополнительной разметки. Wav2vec 2.0 имеет архитектуру, представленную на Рис. 1.

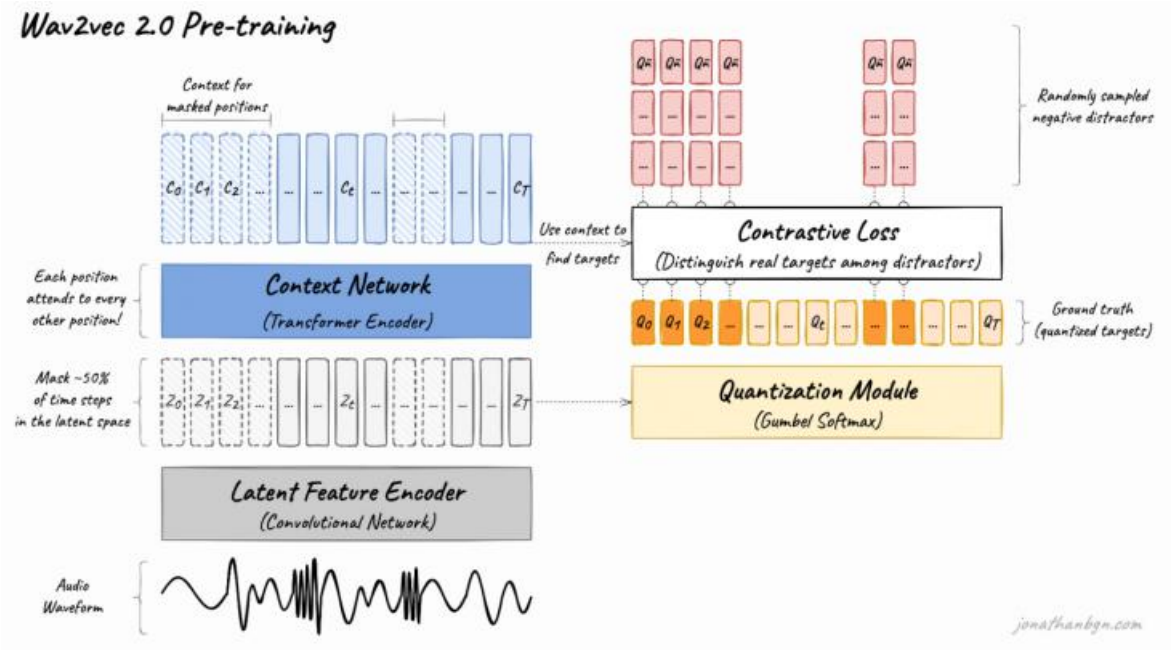

Рис. 1. Архитектура Wav2vec 2.0

Сначала исходная аудиозапись через модуль кодирования исходных данных, который состоит из 7 свёрточных слоёв (Рис. 2) и превращает аудиозапись в последовательность векторов для каждых 20 мс записи. Каждый свёрточный слой имеет размерность 512 (количество фильтров), а размерность фильтра и ширина шага свёртки постепенно снижается с последующими слоями.

Полученные на данном этапе векторы проходят этап квантизации для того, чтобы перевести непрерывные значения в определённый набор конечных. Для этого Wav2vec 2.0 использует две "кодовые группы" (codebooks) из 320 "кодовых слов" (codewords) каждая. Имеющиеся вектора умножаются на матрицу квантизации Wq, после чего к полученному результату применяется функция Gumbel softmax [24], которая выводит для каждого исходного вектора наиболее подходящие для него кодовые слова (по одному из каждой группы). Данные кодовые слова конкатенируются и путём линейной трансформации получается финальная репрезентация.

Следующие этапы используются для предварительного обучения модели. Извлечённые из аудиозаписи вектора (полученные до квантизации) маскируются с вероятностью 50%. Для этого из упорядоченных векторов случайным образом отбираются 0.065% плюс девять им последующих векторов с учётом наложения. Данные вектора заменяются на один и тот же маскирующий вектор. Далее все имеющиеся вектора проходят через проективный слой, который увеличивает их размерность (768 для модели BASE и 1024 для модели LARGE).

Следующий этап является основой модели и представляет из себя трансформер последовательностью из 12 или 24 энкодеров для BASE и LARGE моделей соответственно. Предварительно каждый вектор также проходит слой позиционного кодирования. В отличие от оригинальной модели трансформера, в которой позиционное кодирование реализовано путём конкатенации заранее сгенерированных позиционных

эмбеддингов и входных векторов, в Wav2vec 2.0 используется слой групповой свёртки [25].

Наконец, полученные после прохождения трансформера вектора проходят проективный слой, который уменьшает их размерность до размерности векторов, полученных на этапе квантизации, после чего сравнивает их и высчитывает функцию потерь. Функция потерь высчитывается как сумма контрастивной (contrastive loss) и различительной (diversity loss) ошибок. Для расчёта контрастивной ошибки модель должна на основе полученного вектора предсказать истинный квантизированный вектор Qp из K+1 векторов, где K=100 отвлекающих векторов, извлечённых из других позиций той же записи. Различительная же ошибка используется для регуляризации, чтобы все возможные кодовые слова из групп 320*320 использовались с одинаковой вероятностью.

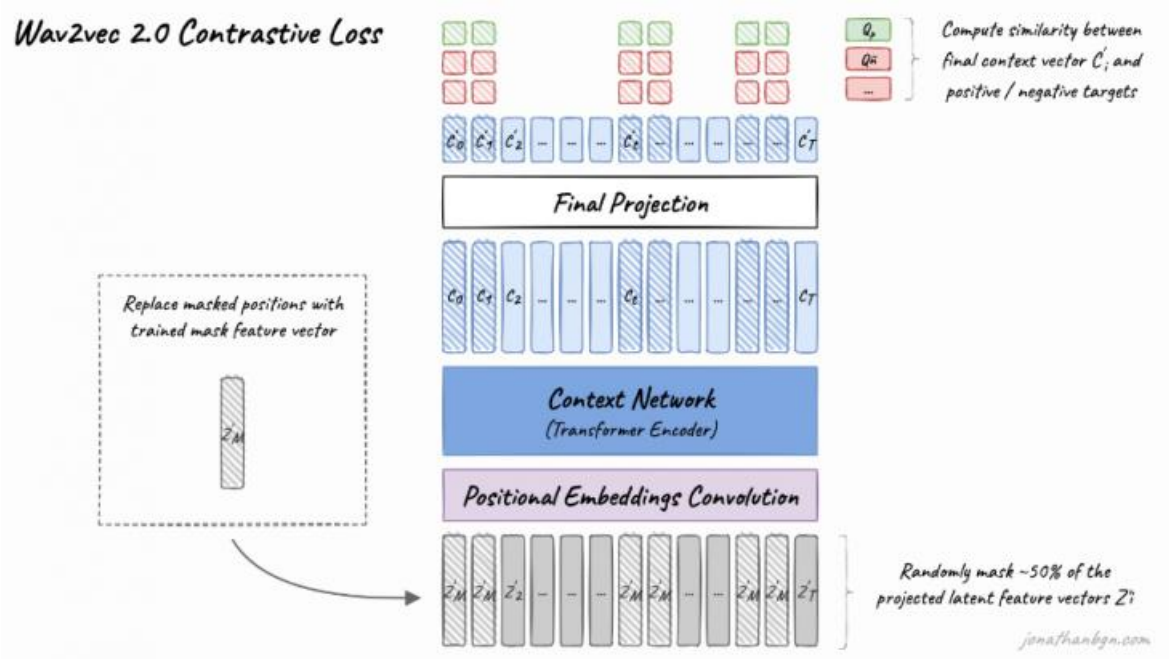

Рис. 2. Процесс предобучения Wav2vec 2.0

Наконец, полученная модель готова для дообучения на соответствующей задаче.

HuBERT использует схожий подход, но при обучении решается задача классификации, в ходе которой модель предсказывает каждому фрейму один из конечного списка классов. Алгоритм состоит из двух основных этапов: этапа создания классов и этапа предсказания.

На этапе создания классов HuBERT извлекает непосредственно из аудиозаписи MFCC-признаки для каждых 20 мс записи, после чего полученные вектора кластеризуются с помощью метода k-средних. После этого каждому вектору присваивается соответствующий класс, и для каждого класса создаётся эмбеддинг, который затем используется при вычислении функции потерь.

По архитектуре этап предсказания модели HuBERT полностью повторяет модель Wav2vec 2.0 за исключением двух вещей: во-первых, предсказываются, как уже было указано, не прошедшие квантизацию вектора, а классы замаскированных векторов, а во-вторых, в HuBERT в качестве функции потерь используется функция кросс-энтропии. Исходные признаки, точно так же как и в Wav2vec 2.0, извлекаются с помощью свёрточной нейронной сети.

После первой итерации HuBERT обновляет свой набор классов для предсказания. Для этого на второй итерации используются вектора, полученные на 6 слое энкодеров, которые точно так же кластеризуются методом k-средних (Рис. 3). Модель BASE имеет всего 2 итерации, модели LARGE и X-LARGE имеют уже 3 итерации. На третьей итерации используются уже векторы, полученные на 9 слое энкодеров второй итерации. Модель X-LARGE отличается от модели LARGE удвоенным количество слоёв энкодеров.

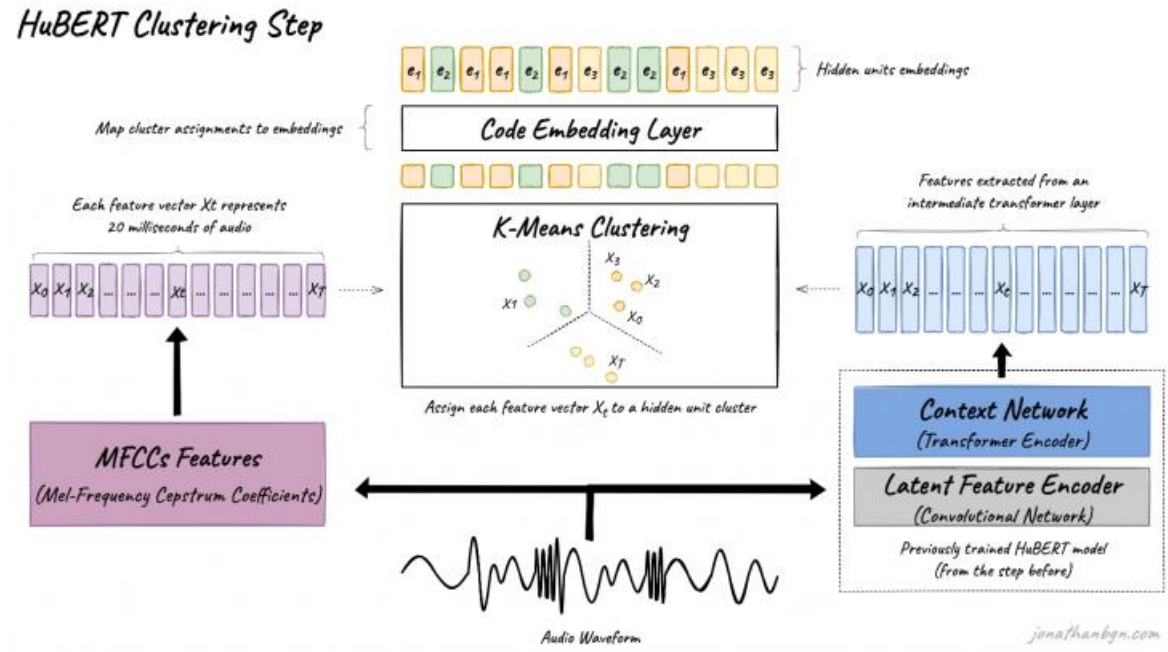

Рис. 3. Этап кластеризации модели HuBERT

WavLM использует тот же подход предсказания псевдо-классов для фреймов как и HuBERT, но также ещё добавляет задачу разделения говорящих и шумоподавления. Для этого на замаскированные вектора накладывается шум или фрагмент аудиозаписи, и модели нужно предсказать класс исходного вектора. Тем самым модель обучается не только задаче распознавания речи, но и, например, её разделению на разных говорящих. Архитектура WavLM также повторяет HuBERT, но в ней произведена небольшая модификация: если в HuBERT используется обычное относительное позиционное кодирование, то в WavLM используется относительное позиционное кодирование с гейтами. Идея гейтов позаимствована из рекуррентной нейросетевой архитектуры GRU [26] и позволяет учитывать содержание речевых фрагментов при кодировании расстояния между ними. WavLM на данный момент показывает наивысшее качество среди всех моделей Self-Supervised Learning в задаче распознавания эмоций на датасете IEMOCAP – 70.62% accuracy (на основе бенчмарка SUPERB [27]).

4. Тестирование

Для исследования эффективности применения описанных методов для автоматического распознавания эмоций в русском языке было проведено их тестирование на русских датасетах Dusha и RESD. Для испытаний были взяты модели wav2vec 2.0, HuBERT и WavLM. Конкретно использовались лежащие в свободном доступе модели на платформе Hugging Face, предварительно обученные на датасете Dusha, и модели из библиотеки Animore, предварительно обученные на датасете RESD. При тестировании все модели прошли дообучение на этом самом датасете с параметрами: 2 эпохи, размер батча 32 (2 * 16 шагов накопления градиента), AdamW в качестве оптимизатора. Результаты представлены в Таблицах 2 и 3 в формате "название_модели —

название_изначального_датасета". При этом у моделей из библиотеки Aniemore представлены те результаты, которые указаны авторами изначальной модели.

В качестве основной метрики использовалась Accuracy (также встречается название Weighted Accuracy (WA)), которая считается стандартной в задаче автоматического распознавания эмоций и вычисляется как соотношение правильно предсказанных классов к неправильно предсказанным.

Модель	Гнев	Позитив	Грусть	Нейтральная	Другое	Accuracy
HuBERT_Base-Dusha	0.83	0.78	0.79	0.94	0.71	0.866
HuBERT_Large-Dusha	0.87	0.81	0.80	0.93	0.7	0.8745
WavLM_Base-Dusha	0.87	0.77	0.83	0.94	0.82	0.872
WavLM_Large-Dusha	0.86	0.81	0.84	0.93	0.75	0.8782
Wav2vec2-RESD	0.66	0.59	0.59	0.9	0.59	0.8063
HuBERT-RESD	0.68	0.59	0.49	0.91	0.64	0.8134
WavLM-RESD	0.65	0.53	0.54	0.92	0.63	0.8015

Таблица 2. Обучение и тестирование моделей на датасете Dusha

Модель	Гнев	Радость	Энтузиазм	Грусть	Отвращение	Страх	Нейтральная
HuBERT_Base-Dusha	0.34	0	0.03	0.03	0.22	0.02	0.39
HuBERT_Large-Dusha	0.14	0.48	0.1	0	0.03	0.33	0.11
WavLM_Base-Dusha	0.73	0.11	0	0.14	0.03	0.04	0
WavLM_Large-Dusha	0.11	0	0	0.06	0.32	0	0.42
Wav2vec2_Large-RESD	0.91	0.59	0.7	0.63	0.76	0.67	0.79
HuBERT_Large-RESD	0.84	0.7	0.8	0.78	0.73	0.67	0.74
WavLM_Large-RESD	0.89	0.84	0.86	0.84	0.84	0.69	0.76

Таблица 3. Обучение и тестирование моделей на датасете RESD

Обучение проводилось только на аудиоданных, без использования текстовых. При этом среди моделей, предобученных на датасете Dusha, нет моделей wav2vec 2.0. Из имеющихся наилучший результат показывает модель WavLM_Large, предобученная и

протестированная на датасете Dusha. На втором месте с небольшим отставанием HuBERT-Large_Dusha. Подтверждается информация из бенчмарка SUPERB, в том числе то, что модель WavLM_Base показывает себя хуже, чем HuBERT_Large, поэтому количество обучаемых параметров по-прежнему играет важную роль даже при модификации архитектуры.

Перенос моделей с одного датасета на другой показывает худшие результаты, чем непосредственно изначальное обучение. При этом результаты на датасете RESD оказались крайне низкими, что можно объяснить его малым размером.

Что касается эмоций, то здесь моделям в датасете Dusha проще всего определять нейтральную эмоцию и гнев. Хуже всего определяются эмоции из категории "Другое". Возможно, при более подробной классификации данной категории качество результатов также можно было бы улучшить. В датасете RESD же модель WavLM со сравнительно одинаковой точностью определяет все эмоции, кроме страха и нейтральной, которые здесь выделяются хуже. Гнев по-прежнему определяется лучше других эмоций.

5. Заключение

В данной статье были рассмотрены основные подходы в области автоматического распознавания эмоций в речи, использующиеся в данный момент. Наиболее популярным подходом в течение последних лет остаётся подход на основе самообучающихся моделей трансформеров, ориентированных на обработку разговорной речи, таких как Wav2vec 2.0, HuBERT и WavLM.

Этап сбора обучающих данных по-прежнему остаётся критически важным для хороших результатов работы моделей. Относительно недавнее появление открытых готовых русских датасетов, созданных для задач анализа эмоциональной разговорной речи, таких как Dusha и RESD, помогло снизить затраты на данном этапе и дало новые возможности для развития данной области применительно к русскому языку, в частности для независимых исследователей. Особенно хороший результат модели-трансформеры показывают на датасете Dusha, в первую очередь за счёт большого объёма собранных данных.

В качестве дальнейших исследований представляется перспективным изучение возможности дополнительного усложнения моделей анализа речи, например, за счёт учёта выделения разных уровней в речевом сигнале (фонем, слогов, слов) [28], а также использование мультимодального подхода с добавлением письменных данных, дублирующих речь.

Библиография

1. Schneider, S., Alexei Baevski, Ronan Collobert, Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition // ArXiv (Cornell University). 2019.
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G. Emotion recognition in human-computer interaction // IEEE Signal Processing Magazine. 2001. V. 18. No. 1. Pp. 32–80.
3. Kondratenko, V., Sokolov, A., Karpov, N., Kutuzov, O., Savushkin, N., Minkin, F. Large Raw Emotional Dataset with Aggregation Mechanism // ArXiv (Cornell University). 2022.
4. Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S. IEMOCAP: interactive emotional dyadic motion capture database // Language Resources and Evaluation. 2008. V. 42. No. 4, Pp. 335–359.
5. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F.,

Schuller, B. W. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. V. 45. No. 9. Pp. 10745-10759.
6. Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B., Star, K., Hajiyev, E., Pantic, M. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021. V. 43. No. 3. Pp. 1022–1040.
7. Mohamad Nezami, O., Jamshid Lou, P., Karami, M ShEMO: a large-scale validated database for Persian speech emotion detection // Language Resources and Evaluation. 2018. V. 53. No. 3. Pp. 1–16.
8. Inger Samsø Engberg, Anya Varnich Hansen, Ove Kjeld Andersen, Dalsgaard, P. Design, recording and verification of a danish emotional speech database // EUROSPEECH. 1997. V. 4. Pp. 1695-1698.
9. Hozjan, V., Kačič, Z. Context-Independent Multilingual Emotion Recognition from Speech Signals // International Journal of Speech Technology. 2003. V. 6. Pp. 311–320.
10. Lotfian, R., Busso, C. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings // IEEE Transactions on Affective Computing. 2019. V. 10. No. 4. Pp. 471–483.
11. Grimm, M., Kroschel, K., Narayanan, S. The Vera am Mittag German audio-visual emotional speech database // International Conference on Multimedia and Expo. 2008. Pp. 865-868.
12. 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. V. 13. No. 5.
13. Lubenets, I., Davidchuk, N., Amentes, A. Aniemore. GitHub. 2022. URL: https://github.com/aniemore/Aniemore
14. Andrew, A. M. An Introduction to Support Vector Machines and Other Kernel‐based Learning Methods // Kybernetes. 2001. V. 30. No. 1. Pp. 103–115.
15. Ho, T. K. Random decision forests // Proceedings of 3rd international conference on document analysis and recognition. 1995. V. 1. Pp. 278–282.
16. Ali, S., Tanweer, S., Khalid, S., Rao, N. Mel Frequency Cepstral Coefficient: A Review // ICIDSSD. 2021.
17. Zheng, W. Q., Yu, J. S., Zou, Y. X. An experimental study of speech emotion recognition based on deep convolutional neural networks // 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). 2015. Pp. 827-831.
18. Hochreiter, S., Schmidhuber, J. Long short-term memory // Neural computation. 1997. V. 9. No. 8. Pp. 1735-1780.
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention Is All You Need // ArXiv (Cornell University). 2017.
20. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // ArXiv (Cornell University). 2018.
21. Baevski, A., Zhou, H., Mohamed, A., Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // ArXiv (Cornell University). 2020.
22. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. V. 29. Pp. 3451–3460.
23. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing // IEEE

Journal of Selected Topics in Signal Processing. 2022. V. 16. No. 6. Pp. 1505–1518.
24. Jang, E., Gu, S., Poole, B. Categorical Reparametrization with Gumbel-Softmax // ArXiv (Cornell University). 2016.
25. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks // Communications of the ACM. 2012. V. 60. No.6. Pp. 84–90.
26. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation // ArXiv (Cornell University). 2014.
27. Yang, S., Chi, P. H., Chuang, Y. S., Lai, C. I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G. T., Huang, T. H., Tseng, W. C., Lee, K., Liu, D. R., Huang, Z., Dong, S., Li, S. W., Watanabe, S., Mohamed, A., Lee, H. SUPERB: Speech processing Universal PERformance Benchmark // ArXiv (Cornell University). 2021.
28. Chen, W., Xing, X., Xu, X., Pang, J., Du, L. SpeechFormer++: A Hierarchical Efficient Framework for Paralinguistic Speech Processing // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023. V. 31. Pp. 775–788

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Представленная на рассмотрение статья «Автоматическая классификация эмоций в речи: методы и данные», предлагаемая к публикации в журнале «Litera», несомненно, является актуальной, ввиду обращения автора к исследованию эмоций с применением техническим средств.
Цель статьи – сравнить между собой результаты часто применяемых методов на основе транформеров на материале русского языка.
В данной статье были рассмотрены основные подходы в области автоматического распознавания эмоций в речи, использующиеся в данный момент.
В разделе 1 дано описание инвентарей эмоций, на основе которых производится классификация. Раздел 2 посвящен сбору и обработке данных, а также имеющимся на данный момент корпусам эмоциональной речи. В разделе 3 приведены методы классификации, которые активно применяются исследователями в своих работах, в том числе и в этой статье. Наконец, в разделе 4 описано тестирование и сравнение результатов классификации эмоций, полученных разными моделями.
Данная работа выполнена профессионально, с соблюдением основных канонов научного исследования. Отметим скрупулёзный труд автора по отбору практического материала и его анализу. Статья является новаторской, одной из первых в российской лингвистике, посвященной исследованию подобной тематики в 21 веке. В статье представлена методология исследования, выбор которой вполне адекватен целям и задачам работы. Автор обращается, в том числе, к различным методам для подтверждения выдвинутой гипотезы.
Для решения исследовательских задач в статье использовались как общенаучные методы, так лингвистические методы. Рассмотрение проблемы взаимодействия искусственного интеллекта и естественной языковой системы осуществляется с помощью методов лингвистического наблюдения и описания, также, применяются общефилософские методы анализа и синтеза.
Исследование выполнено в русле современных научных подходов, работа состоит из введения, содержащего постановку проблемы, основной части, традиционно

начинающуюся с обзора теоретических источников и научных направлений, исследовательскую и заключительную, в которой представлены выводы, полученные автором. Библиография статьи насчитывает 28 источников, среди которых представлены труды зарубежных исследователей на иностранном языке. Считаем, что обращение к работам российских исследователей усилило бы теоретическую составляющую работы.

К сожалению, в статье отсутствуют ссылки на фундаментальные работы, такие как кандидатские и докторские диссертации.

В общем и целом, следует отметить, что статья написана простым, понятным для читателя языком. Опечатки, грамматические и стилистические ошибки не выявлены.

Данные, полученные в результате исследования, представлены в формате графиков и диаграмм, что облегчает восприятие читателем.

Работа является новаторской, представляющей авторское видение решения рассматриваемого вопроса. Статья, несомненно, будет полезна широкому кругу лиц, филологам, магистрантам и аспирантам профильных вузов. Практическая значимость исследования определяется возможностью применения данных статьи в курсах по теории языка и компьютерной лингвистике. Статья «Автоматическая классификация эмоций в речи: методы и данные» может быть рекомендована к публикации в научном журнале.