

Межязыковой перенос моделей автоматического распознавания эмоций в речи на русский язык: данные и тенденции

Представлено описание исследования влияния различий в языках и обучающих данных на качество межязыкового переноса обученной речевой модели на русский язык в задаче автоматического распознавания эмоций в речи. На этапе обучения в качестве языков-источников выступили английский, польский, китайский и японский языки, для которых соответственно использовались датасеты эмоциональной речи IEMOCAP, nEMO, ESD и JNVN, а самой моделью – речевая модель HuBERT на архитектуре трансформера. Все обученные на соответствующем датасете модели протестированы на укороченной выборке из датасета русской эмоциональной речи Dusha. На основе полученных данных рассмотрены основные тенденции при выборе разных языков для обучения речевой модели и её дальнейшего переноса на русский язык, а также проанализированы различия в датасетах, которые указывают на необходимость дальнейшей работы по сбору и разметке качественных данных эмоциональной речи.

Ключевые слова: обработка естественного языка, машинное обучение, трансформеры, распознавание эмоций, распознавание речи, межязыковой перенос моделей, HuBERT, Dusha, IEMOCAP, nEMO, ESD, JNVN

DOI: 10.36535/0548-0027-2025-05-4

ВВЕДЕНИЕ

Автоматическое распознавание эмоций в разговорной речи представляет собой одну из фундаментальных областей обработки естественного языка, которая затрагивает невербальные методы коммуникации между людьми. Распознавание внутреннего эмоционального состояния человека является существенным для взаимодействия человека и компьютера, например, для постановки диагноза в области здравоохранения, [1] или для обеспечения общественной безопасности [2]. При классификации обычно используется набор небольшого количества архетипических эмоций, универсальных для всех языков, либо двумерное пространство «валентность/возбуждение», на котором модель предсказывает точку, соответствующую определённому высказыванию. «Возбуждение» на таком пространстве отражает степень проявления эмоции, тогда как «валентность» – определяет положительная или отрицательная эта эмоция.

Несмотря на значительные успехи в задаче классификации эмоций в рамках одного датасета, перенос речевых моделей – алгоритмов, обученных на определённом наборе данных и способных затем самостоятельно определять эмоции, выраженные в речи –

на новые области или новые языки пока ещё не достиг тех же результатов, на которые способен человек [3]. Малое количество размеченных данных для обучения моделей по-прежнему остаётся главной проблемой для многих языков. Потенциальным представляется решение использовать данные других языков для обучения речевой модели с последующим переносом ее параметров на целевой язык. Несмотря на то, что имеется значительное количество работ на тему межязыкового переноса моделей, обученных на текстовых данных, о возможностях переноса моделей, обученных на эмоционально окрашенных речевых данных, в частности, современных моделей с архитектурой трансформера, исследований было проведено сравнительно мало. Задачу межязыкового переноса часто осложняют различия в подходах к составлению баз данных, в том числе в наборе используемых эмоций, и проблема, получившая название «проклятие мульти-модальности» [4], заключающаяся в том, что есть предел языков, которые модель может покрыть до того момента, как начнёт снижаться её качество.

При межязыковом переносе речевых моделей выделяют два возможных варианта их применения: с дообучением на небольшой выборке (few-shot) при-

меров из целевого языка или без дополнительного дообучения (zero-shot) на целевом языке совсем [5]. Есть публикации [4, 6], в которых авторы применяли автоматический перевод предложений с целевого языка для дополнительного обогащения модели обучающими данными.

Цель настоящей статьи – исследование качества переноса моделей на материале русского языка, при этом не ставилась задача изучения современных методов машинного обучения в области автоматического распознавания эмоций в речи, а лишь обращается внимание на различие между самими языками и имеющимися для них данными для обучения. С этой целью было проведено обучение речевой модели с архитектурой трансформера HuBERT [7] на разных наборах данных иностранных языков с задачей распознавания эмоции в речи, после чего модель была перенесена и протестирована на данных русского языка. Данными для русского языка послужил датасет Dusha [8] – крупнейший русский датасет эмоциональной речи на данный момент. В качестве иностранных языков для обучения и дальнейшего переноса модели были выбраны английский, польский, китайский и японский языки из-за наличия для них качественно собранных датасетов эмоциональной речи. Исследование придерживается подхода zero-shot при переносе модели, тем самым при её обучении не были использованы данные целевого (русского) языка.

КРАТКИЙ ОБЗОР ПУБЛИКАЦИЙ ПО ТЕМЕ

Большая часть уже имеющихся исследований о влиянии языковых признаков на качество результатов при межязыковом переносе обученных моделей сфокусирована на текстовых данных. В обзорной работе [9] авторы выделяют пять основных факторов, влияющих на качество переноса: лингвистическое сходство; пересечение лексики; обучающие данные; архитектура модели; параметры обучения.

Так в настоящей работе рассматривается в первую очередь влияние на качество переноса модели выбор самих языков и обучающих данных, то мы сосредоточимся только на первых трёх факторах.

Для определения *лингвистического сходства* языков обычно применяются база данных WALS [10], содержащая типологическую классификацию лингвистических признаков, и векторы lang2vec, используемые для оценки косинусной близости между языками на основе данных типологической базы URIEL [11].

Синтаксис и порядок слов принято считать наиболее важными признаками, влияющими на качество переноса модели, однако при более близком рассмотрении они могут оказывать менее значимое влияние. Так, авторы работ [12] и [13] провели сравнения разных порядков слов в рамках одного языка, в частности, случайно переставляя слова и полноценно адаптируя порядок слов к правилам другого языка. Как оказалось, случайные перестановки снижают качество модели значительно больше, чем адаптация под реальный порядок слов другого языка.

Среди лингвистических признаков, оказывающих влияние на качество переноса модели, часто выделяют *географическую, генетическую и фонологическую*

близость [14]. Для задачи распознавания эмоций в речи, по сравнению с текстами, наибольший интерес представляет фонологическая близость и разница в фонологическом инвентаре языков, однако, к сожалению, исследовательских работ на эту тему на данный момент имеется мало. Одна из них – это статья [15], в которой авторы описали схожесть в значениях формант некоторых гласных в разных языках при выражении эмоций, на основе чего затем была построена модель распознавания эмоций в речи со значительно более высоким (приблизительно на 7%) качеством классификации при её переносе на другой язык по сравнению с базовой моделью. Влияние генетической близости языков при распознавании эмоций также показано в работе [16], результаты которой обращают внимание на улучшение качества классификации при переносе моделей внутри одной языковой семьи, по сравнению с переносом на языки других языковых семей.

Второй важный фактор, влияющий на качество переноса модели – это *пересечение лексики*. Оно оценивается как доля слов, присутствующих в обоих исследуемых языках. Его можно посчитать, например, на основе двух одноязычных корпусов, поделив количество присутствующих в них слов на общее количество уникальных слов. Среди других метрик часто используются ezGlot [17] и нормализованное расстояние Левенштейна [18].

Среди исследований на эту тему особенно следует выделить работу [13]. В ней авторы использовали перенос модели с синтетического английского языка на стандартный английский, благодаря чему они изолировали влияние пересечения лексики от остальных переменных и смогли изучать его взаимодействие с другими лингвистическими признаками. На основе экспериментов авторы сделали вывод, что пересечение лексики оказывает наибольшее влияние на качество переноса модели тогда, когда порядок слов в языках различается. В работе [19] исследовалось влияние пересечения лексики на качество переноса. В результате было выявлено, что качество переноса сильнее зависит от количества пересекающейся лексики в случае малого количества исходных обучающих данных. В целом стоит сказать, что пересечение лексики не является каким-то независимым фактором и нужно смотреть на его сочетание с другими факторами.

Что касается *обучающих данных*, то в качестве двух основных факторов, влияющих на качество переноса модели, здесь выделяют *размер обучающего корпуса* и *соответствие между собой обучающих данных мультиязычной модели*. Так, влияние размера корпуса на качество переноса продемонстрировано в [20], где авторы сравнили качество работы текстовой мультиязычной модели BERT при обучении на корпусах разного размера и зафиксировали существенный рост качества в случае большего объёма обучающего корпуса. Если рассматривать соответствие между обучающими данными, то здесь подразумевается, во-первых, параллельность корпусов, на которых обучается мультиязычная модель, а, во-вторых, принадлежность данных к одной и той же области знания. Причём, как показано в [13], качество переноса мультиязычной модели значительно ухудшается

в случае обучения на данных, принадлежащих к разным областям, чем на данных, принадлежащих к одной области, но не из параллельных корпусов.

ДАТАСЕТЫ И ЯЗЫКИ

В качестве данных для обучения в настоящей работе были взяты датасеты IEMOCAP [21], nEMO [22], ESD [23] и JNVN [24], содержащие эмоциональную речь на английском, польском, китайском и японском языках соответственно. На их основе было проведено обучение моделей для задачи автоматического распознавания эмоций с целью дальнейшего переноса и тестирования этих моделей на данных русского датасета эмоциональной речи Dusha. Кроме того, из-за разницы в количестве примеров и эмоций между разными датасетами, для каждого языка была создана нормализованная укороченная версия соответствующего датасета в 600 высказываний. Для этого из каждого датасета было случайным образом выбрано по 200 высказываний по каждой из трёх эмоций: Гнев, Радость и Грусть. Эти эмоции были выбраны на основе датасета Dusha, исключая из него нейтральное состояние и категорию «Другое» (категория «Позитив» соответствует эмоции Радость в данном исследовании).

Из этих языков наиболее сходными с русским – на основе признаков, описанных в предыдущем разделе, – являются английский и польский языки, в противовес китайскому и японскому. В табл. 1 приведены различия русского и других языков по основным лингвистическим признакам, рассчитанные на основе векторов lang2vec. Подробные сведения о признаках и сходстве языков приведены в *Приложении*.

Dusha был 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 на основе общения реальных людей с виртуальным ассистентом сгенерировали тексты, которые затем были озвучены с помощью краудсорсинга (озвучивающим людям давался текст и эмоция, с которой этот текст должен быть произнесён, затем полученные аудиозаписи дополнительно проверяла вторая группа разметчиков); 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, а также их письменные расшифровки.

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) – мультимодальный датасет, созданный для исследования коммуникации между людьми и состоящий из более 151 часа записанных разговоров между актёрами, поделенных на сегменты около 4,5 секунд каждый. 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 9 эмоций: Гнев, Радость, Грусть, Отвращение, Страх, Удивление, Возбуждение, Раздражение, Разочарование и Нейтральное состояние. Этот датасет в настоящее время является наиболее популярным в задачах автоматического распознавания эмоций.

nEMO – корпус польской эмоциональной речи, содержит 4481 высказывание общей длительностью более 3-х часов. Содержание высказываний было сформировано с целью минимизировать их количество, отразив при этом все имеющиеся в польском языке звуки речи. В качестве говорящих выступили 9 носителей польского языка в возрасте 20-30 лет, три из которых являются профессиональными актёрами.

ESD – параллельный корпус эмоциональной речи для китайского и английского языков, содержит 35000 высказываний, записанных при помощи 20 говорящих – 10 на китайском и 10 на английском языках. Всего в корпусе содержится 350 уникальных по содержанию высказываний, т.е. каждый говорящий записывает по 350 высказываний для каждой из пяти эмоций, представленных в корпусе: Гнев, Радость, Грусть, Удивление и Нейтральное состояние. Для снижения вариативности экстралингвистических факторов возраст каждого говорящего составляет от 25 до 35 лет. Корпус создавался в первую очередь для задач по автоматическому переносу эмоций в речи, когда нужно изменить эмоцию высказывания, сохранив его лексический состав. В настоящей работе использовались только аудиозаписи на китайском языке, т. е. половина корпуса.

Таблица 1

Различия русского с другими языками по основным лингвистическим признакам.
Значения основаны на векторах lang2vec, чем значение ближе к нулю, тем выше сходство языков

Различия	Язык			
	английский	польский	китайский	японский
синтаксическое	0,49	0,47	0,57	0,66
фонологическое	0,28	0	0,62	0,43
генетическое	0,9	0,4	1	1
географическая дистанция	0,2	0,2	1	0,3
фонемного инвентаря	0,56	0,49	0,69	0,54

Данные о датасетах, использованных в настоящей работе

Датасет	Объем	Количество говорящих	Метод создания	Язык	Эмоции
Dusha	300000 записей (350 часов)	Более 2000	Прочтение реплик / извлечение из подкастов	Русский	Гнев, Позитив, Грусть, Другое, Нейтральное состояние
IEMOCAP	10039 высказываний	10	Прочтение реплик и импровизация (диалог)	Английский	Гнев, Радость, Грусть, Отвращение, Страх, Удивление, Волнение, Раздражение, Разочарование, Нейтральное состояние
nEMO	4481 высказываний (более 3 часов)	9	0,4	Польский	Гнев, Радость, Грусть, Страх, Удивление, Нейтральное состояние
ESD	35000 высказываний	20	0,2	Китайский, английский	Гнев, Радость, Грусть, Удивление, Нейтральное состояние
JVNV	1615 высказываний (около 4 часов)	4	0,49	Японский	Гнев, Радость, Грусть, Отвращение, Страх, Удивление

JVNV – корпус японской эмоциональной речи, который имеет две важные особенности: во-первых, каждое высказывание содержит как минимум одно невербальное речевое выражение (nonverbal vocalization), например, смех, а, во-вторых, тексты всех высказываний, которые затем были озвучены профессиональными актёрами, были автоматически сгенерированы с помощью больших языковых моделей (large language model, LLM). Для этого авторы делали в LLM запрос о генерации новых предложений на основе эмоционально окрашенных словарных слов и невербальных речевых выражений, после чего вручную отбирали наиболее качественные результаты. Такой подход позволил значительно снизить затраты на сбор данных. Датасет JVNV содержит 1615 высказываний общей длительностью около 4-х часов и 6 базовых эмоций: Гнев, Радость, Грусть, Отвращение, Страх и Удивление. Все высказывания были зачитаны профессиональными актёрами (2 женщины и 2 мужчины). В табл. 2 приведены обобщенные данные описанных датасетов.

МОДЕЛИ И ПАРАМЕТРЫ

В качестве основной архитектуры, использованной для исследования качества результатов при меязыковом переносе моделей, была принята архитектура HuBERT-base [7], содержащая около 95 млн параметров. При обучении она выполняет задачу классификации: маскирует часть поступающих на вход «фреймов» – коротких отрезков аудиозаписи, длительностью обычно 20 мс, – которые затем учится классифицировать на основе конечного набора классов. Этот набор классов, в свою очередь, формируется с помощью применения к незамаскированным фреймам метода k-средних. Для задачи классификации высказывания по эмоциям к этой модели на конце добавлен слой классификатора.

Всего для каждого языка было протестировано 5 моделей. Все они тестировались на укороченном датасете Dusha из 600 высказываний. Для первой модели (base) архитектура HuBERT была обучена на соответствующем иностранном датасете из 600 высказываний (20% датасета было использовано в качестве валидационной выборки). Цель второй модели (full) – сравнение качества переноса при обучении модели на полном датасете и его небольшой части, поэтому для второй модели при обучении HuBERT использовали уже не укороченные, а полные датасеты каждого языка, после чего для классификации только трёх эмоций была проведена настройка слоя классификатора модели на соответствующем укороченном датасете. Цель третьей модели (pretrain_freeze) – исследование влияния данных других языков, которые могли быть использованы при обучении модели: для этого было произведено дообучение модели Hubert-base-ls960 [7], которая была предобучена на данных корпуса английской речи LibriSpeech длительностью около 960 часов, при этом само дообучение идентично обучению второй модели. Четвёртая модель (pretrain) повторяет предыдущую, но с одной разницей: при дообучении третьей модели была произведена заморозка всех параметров, кроме последних двух слоёв, тогда как при дообучении четвёртой модели заморозка параметров не производилась совсем. Заморозка в данном случае может помочь сохранить параметры исходной модели без изменений, которые на первых уровнях, как считается, обычно моделируют более локальные признаки (звуки и слова), когда на более высоких слоях модели параметры лучше моделируют уже непосредственно глобальные признаки конкретной задачи (в данном случае – распознавания эмоций). Наконец, в качестве последней модели использовалась мультязычная модель mHuBERT-147 [25], предобученная на аудиоданных 147 языков, включающих в себя в том числе

русский, английский, польский, китайский и японский языки длительностью более 90 тыс. часов. Для каждого языка мультязычная модель прошла дообучение на соответствующем укороченном датасете из 600 высказываний без заморозки параметров.

РЕЗУЛЬТАТЫ

Представленные в настоящей статье речевые модели, были обучены на иностранных датасетах IEMOCAP, nEMO, ESD и JVNv и их укороченных версиях из 3-х эмоций и 600 высказываний, в общей сложности по пять моделей на одном иностранном языке, и протестированы на укороченном датасете Dusha из 600 высказываний по 200 на каждую из трёх эмоций (Гнев, Радость и Грусть). Для каждого иностранного языка представлены пять моделей:

- base – модель, обученная с нуля на укороченном датасете;
- full – модель, обученная с нуля на полном датасете;
- pretrain_freeze и pretrain – предобученные модели Hubert-base-ls960, дообученные на соответствующем полном датасете с заморозкой слоёв и без соответственно;
- multi – мультязычная модель.

Общее качество классификаций каждой модели было подсчитано на основе метрики Unweighted Accuracy (UA), отражающей отношение количества правильно классифицированных примеров к неправильным. Для каждой отдельной эмоции было подсчитано качество её классификации на основе F-меры. В табл. 3 приведены полученные результаты для каждой модели. Для более детального анализа мы также включили в неё процентное распределение классифицированных

высказываний по эмоциям (для каждой эмоции соответствующий столбец указан как count). Перед названием каждой модели указан префикс в виде первой буквы названия датасета, на данных которого обучалась модель – IEMOCAP, nEMO, ESD и JVNv в соответствующем порядке.

Обсуждение результатов начнём с модели base. Можно увидеть, что общее качество переноса моделей для английского, польского и японского языков в целом сходно, при этом наилучшее качество наблюдается у польской и китайской моделей N-base и E-base соответственно. Классификация различных эмоций тоже имеет схожую тенденцию: лучше всего классифицируются эмоции Гнев и Грусть, в то время как эмоция Радость классифицируется значительно хуже. Это можно объяснить качеством самих эмоций, если расположить их в пространстве «возбуждение/валентность»: Гнев и Радость имеют высокое значение возбуждения, обычно выражаемое в более быстрой, энергичной и громкой речи, в то время как Грусть наоборот имеет низкое значение возбуждения, таким образом позволяя более легко противопоставить её остальным эмоциям. Гнев и Радость, однако, противопоставлены между собой только по шкале валентности, эксплицитно не соотносящейся с акустическими признаками высказывания, которые мог бы использовать классификатор, из-за чего он большую часть таких высказываний относит к эмоции Гнев. Последние исследования указывают, что со шкалой валентности главным образом может соотноситься имплицитная лингвистическая информация, содержащаяся в речи [26]. На общем фоне при этом выделяется китайская модель E-base: у неё, во-первых, наихудшее качество классификации эмоции Радость среди всех моделей base, а, во-вторых, почти две трети всех высказываний ею классифицированы как Грусть.

Таблица 3

Результаты моделей для каждого иностранного датасета*

Модель	Качество классификации						
	anger-f1	anger-count, %	happy-f1	happy-count, %	sad-f1	sad-count, %	UA
I-base	0,46	36,7	0,28	29,5	0,51	33,8	0,42
I-full	0,52	77	0,03	1,2	0,44	21,8	0,41
I-pretrain_freeze	0,58	50,5	0,35	31,8	0,37	17,7	0,45
I-pretrain	0,65	53,2	0,25	18,3	0,57	28,5	0,52
I-multi	0,63	44,2	0,24	11,5	0,73	44,3	0,59
N-base	0,48	45,3	0,33	24,8	0,54	29,8	0,46
N-full	0,38	29,8	0,48	43,3	0,53	26,8	0,46
N-pretrain_freeze	0,54	18,2	0,41	11	0,63	70,8	0,56
N-pretrain	0,60	19,3	0,40	12,3	0,65	68,3	0,58
N-multi	0,58	18,2	0,68	24,2	0,73	57,7	0,68
E-base	0,46	28,5	0,16	10,2	0,60	61,3	0,46
E-full	0,12	4,3	0,35	26,2	0,59	69,5	0,43
E-pretrain_freeze	0,16	5,2	0,41	14,3	0,57	80,5	0,45
E-pretrain	0,26	7,2	0,17	5,5	0,55	87,3	0,42
E-multi	0,24	5,7	0,33	12	0,58	82,3	0,46
J-base	0,47	38,8	0,34	33,3	0,47	27,8	0,43
J-full	0,44	28,7	0,18	12,5	0,56	58,8	0,43
J-pretrain_freeze	0,48	81,3	0,08	1,8	0,26	16,8	0,35
J-pretrain	0,49	49,7	0,11	12,5	0,26	37,8	0,45
J-multi	0,54	37	0,09	2,8	0,67	60,2	0,52

* Полученные результаты для всех примеров также находятся в открытом доступе по ссылке https://huggingface.co/datasets/Klos3/SER_transfer_russian

Увеличение обучающей выборки в моделях full даёт разные результаты их переноса для разных датасетов: по сравнению с моделями N-base и E-base, у польской и китайской моделей N-full и E-full улучшилось качество классификации эмоции Радость и наоборот ухудшилось качество эмоции Гнев, при этом общее качество классификации в случае польского языка сохранилось на том же уровне, а для китайского упало. У английского и японского языков сильно упало качество классификации эмоции Радость, при этом английская модель отнесла более 75% высказываний к эмоции Гнев (и всего лишь 1% к эмоции Радость), а японская модель отнесла более 55% высказываний к эмоции Грусть.

Для того чтобы выдвинуть предположения о полученных результатах, в первую очередь нужно взглянуть на сами датасеты. Датасеты nEMO и ESD сбалансированы по количеству высказываний, при этом оба имеют также эмоцию Удивление, которая в плане выражения во многом сходна с эмоцией Радость (высокое возбуждение и относительно положительная валентность), из-за чего модели могли начать относить больше примеров к эмоции Радость, нежели к эмоции Гнев. С датасетом JNVN немного иная ситуация: во-первых, у него нет нейтрального состояния, а, во-вторых, эмоции Страх, Удивление и Отвращение, присутствующие в полном датасете, имеют высокую степень возбуждения и интенсивное выражение в речи, тем самым эмоция Грусть остаётся в датасете единственной эмоцией с низким возбуждением, из-за чего модель может относить к ней большинство высказываний, если они в тестовом датасете не имеют достаточную интенсивность. Наконец, датасет IEMOCAP не сбалансирован, в нём преобладают высказывания с негативной валентностью: почти треть высказываний имеют эмоцию Раздражение, в то время как все высказывания с позитивной валентностью (Радость, Волнение, Удивление) составляют только четверть всего датасета. Из-за этого модель full, обученная на IEMOCAP, может иметь тенденцию относить большинство высказываний с высоким возбуждением к эмоции Гнев, а с низким – к эмоции Грусть (на это указывает почти полное отсутствие высказываний, отнесённых к эмоции Радость).

Говоря про различия в датасетах, стоит также рассмотреть не только разницу в наборе эмоций, но и методы сбора самих данных. Все иностранные датасеты, записывались в специально предусмотренных условиях с участием профессиональных/полупрофессиональных актёров, и, следовательно, имеют более интенсивное выражение эмоций. Небольшое отличие имеет лишь датасет IEMOCAP, так как высказывания в нём записывались в формате диалога между говорящими, из-за чего они гораздо ближе к бытовому разговору. Русский же датасет Dusha был собран посредством краудсорсинга с привлечение большого количества людей, из-за чего озвученные высказывания и эмоции часто имеют значительное различие между собой. Это выражается и в качестве аудиозаписи (нет единой унификации по громкости, некоторые записи очень тихие или вовсе неразборчивые,

большое количество фонового шума), и в качестве самого прочтения эмоций (некоторые записи прочитаны близким к монотонному голосом, из-за чего эмоцию высказывания можно понять только из лексики и содержания, либо эмоция может неоднозначно восприниматься даже носителями языка). В каком-то смысле Dusha тем самым ближе к реальной разговорной речи по сравнению с другими датасетами, но из-за этого также возникают и значительные сложности переноса – например, распознавание большинства примеров моделью J-full как эмоция Грусть с наибольшей вероятностью вызвано именно большей монотонностью прочтения русских высказываний в датасете Dusha по сравнению с прочтением японских высказываний профессиональными актёрами в датасете JNVN.

Далее рассмотрим модели pretrain и pretrain_freeze, полученные дообучением уже предобученной модели Hubert-base-ls960. Сначала сравним их с моделью full, так как они обучались на одинаковых эмоциональных данных и различаются только наличием дополнительного предобучения на английском языке. У английской и польской моделей pretrain_freeze улучшилось общее качество классификации по сравнению с моделью full. У китайской модели общее качество осталось примерно на схожем уровне, у японской же наоборот снизилось. У всех моделей немного увеличилось качество классификации эмоции Гнев, для других же эмоций качество изменилось по-разному в зависимости от датасета. Сложно сказать, что именно на это повлияло, так как предобученная модель Hubert-base-ls960 не обучалась на задаче классификации эмоций. На отсутствие каких-то значимых эмоциональных данных в параметрах предобученной модели также указывает разное распределение высказываний по классам при дообучении модели на разных датасетах: при наличии данных дополнительного (английского) языка, модели N-pretrain и N-pretrain_freeze классифицируют наибольшую часть высказываний с эмоцией Грусть (по сравнению с моделями N-base и N-full, которые имеют данные только об одном языке), в то время как модели J-pretrain и J-pretrain_freeze классифицируют наибольшую часть высказываний с эмоцией Гнев.

Что касается разницы между самими моделями pretrain и pretrain_freeze, то здесь тенденции не слишком очевидны, так как у разных датасетов получились слишком разные различия по качеству: у английской и польской моделей при разморозке всех параметров наблюдается небольшое улучшение как общего качества классификации, так и качества классификации отдельных эмоций, в то время как у китайской и японской моделей качество классификации при разморозке параметров в основном либо падает, либо остаётся на том же уровне. В целом, в результатах этих моделей на фоне моделей base и full можно проследить следующую тенденцию: при добавлении в модель дополнительных данных из того же языка (модели I-pretrain и I-pretrain_freeze) или относительно не слишком дальнего по признакам языка (модели N-pretrain и N-pretrain_freeze), качество классификации при переносе модели возрастает, при этом его

можно ещё немного улучшить за счёт дополнительной настройки всех параметров модели, а не только самых глубоких слоёв. В то же время, если между языками наблюдается более сильное различие, как между английским и китайским или английским и японским, то добавление дополнительной информации о таких языках особым образом на качество переноса модели не влияет и может даже его ухудшить.

Наконец, посмотрим на результаты работы мультиязычной модели. Даже несмотря на малое количество эмоциональных данных, использованных во время её дообучения (всего 600 примеров для каждого языка), модель multi показывает значительно более высокие результаты для всех языков, за исключением лишь китайского, для которого по качеству она сравнима с моделью base. Улучшение качества классификации по сравнению с моделями pretrain и pretrain-freeze, которые предобучались на большом массиве, но только на одном языке, указывает на значительно большее влияние выбора данных для предобучения (в этом случае – данных сразу нескольких языков, а не только одного) по сравнению с просто увеличением количества обучающих данных.

Отдельно стоит отметить тенденцию всех моделей, обученных на китайском датасете, относить большинство высказываний к эмоции Грусть. К сожалению, у нас нет однозначного предположения о том, почему могли быть получены такие результаты. Все аудиозаписи укороченного датасета ESD были прослушаны вручную и для носителя русского языка они в большинстве случаев однозначно интерпретируются с верной эмоцией даже без знания лексики. Возможно, необходимы дополнительные исследования.

В целом, все имеющиеся результаты указывают в первую очередь на слишком большую разницу в качестве имеющихся датасетов, особенно русского датасета Dusha. К сожалению, для русского языка на данный момент в открытом доступе больше нет датасетов эмоциональной речи схожего объёма и качества, но он уступает датасетам других языков. Для задачи автоматической классификации эмоций в речи, как и для задач на основе письменных текстов, наблюдается схожая тенденция более качественных результатов при переносе модели с одного языка на другой, если язык-источник и целевой язык относительно близки. Однако для задачи автоматической классификации эмоций в речи возникают дополнительные трудности, в первую очередь из-за разных методов создания корпусов и отсутствия фиксированного стандартного набора эмоций, для решения которых необходимы более тщательно и систематично собранные данные. При этом даже с небольшим количеством нормализованных данных можно получить довольно устойчивые результаты межязыкового переноса для данной задачи, как можно увидеть из результатов модели base, обученной на разных языках.

ЗАКЛЮЧЕНИЕ

В настоящей статье исследовано влияние обучающих данных и языков, на которых эти данные были собраны, на качество классификации при межязыко-

вом переносе обученных моделей на русский язык, в частности, для задачи автоматического распознавания эмоций в речи. Эксперимент по переносу моделей, обученных на иностранных данных эмоциональной речи, на русский корпус Dusha показал, что при наличии большей близости языка-источника, на данных которого обучалась модель, и целевого языка, на который происходит сам перенос, качество переноса улучшается. Кроме того, для задачи распознавания эмоций в речи, по сравнению с задачами на письменных данных, наблюдаются дополнительные трудности при переносе моделей, связанные с различием в подходах к сбору данных, что в данном случае наблюдается из-за различия в выборе говорящих – произносящих эмоциональные высказывания. Несмотря на большое количество данных, собранных в русском корпусе Dusha, для улучшения возможностей межязыкового переноса моделей распознавания эмоций в речи на русский язык (или с русского на другие языки) по-прежнему необходима дальнейшая работа по сбору качественно озвученных и размеченных данных.

СПИСОК ЛИТЕРАТУРЫ

1. Hansen L., Zhang Y.P., Wolf D., Sechidis K., Ladegaard N., Fusaroli R. A generalizable speech emotion recognition model reveals depression and remission // *Acta Psychiatrica Scandinavica*. – 2022. – Vol. 145, № 2. – P. 186-199. DOI: 10.1111/acps.13388.
2. Lefter I., Jonker C.M. Aggression recognition using overlapping speech // 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). – San Antonio, TX, USA: IEEE, 2017. – P. 299-304. DOI: 10.1109/ACII.2017.8273616.
3. Han Z., Geng T., Feng H., Yuan J., Richmond K., Li Y. Cross-lingual Speech Emotion Recognition: Humans vs. Self-Supervised Models // *arXiv.org*. – 2024. DOI: 10.48550/arXiv.2409.16920.
4. Conneau A., Lample G. Cross-lingual Language Model Pretraining // *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. – Red Hook, NY, USA: Curran Associates Inc., 2019. – № 634. – P. 7059-7069. DOI: 10.5555/3454287.3454921.
5. Pikuliak M., Šimko M., Bieliková M. Cross-lingual learning for text processing: A survey // *Expert Systems with Applications*. – 2021. – Vol. 165, № 113765. DOI: 10.1016/j.eswa.2020.113765.
6. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E. Unsupervised Cross-lingual Representation Learning at Scale // *Annual Meeting of the Association for Computational Linguistics*. – Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.747.
7. Hsu W.N., Bolte B., Tsai Y.H., Lakhota 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 – IEEE Press. – 2021. – Vol. 29. – P. 3451-3460. DOI: 10.1109/TASLP.2021.3122291.
8. Kondratenko V., Sokolov A., Karpov N., Kutuzov O., Savushkin N., Minkin F. Large Raw Emotional Dataset with Aggregation Mechanism // arXiv.org. – 2022. DOI: 10.48550/arXiv.2212.12266.
 9. Philipp F., Guo S., Haddadan S. Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review // arXiv.org. – 2023. DOI: 10.18653/v1/2023.acl-long.323.
 10. Dryer M.S., Haspelmath M. WALS Online // Zenodo. – 2013. DOI: 10.5281/zenodo.13950591.
 11. Littell P., Mortensen D.R., Lin K., Kairis K., Turner C., Lori Levin L. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. – Valencia, Spain: Association for Computational Linguistics, 2017. – Vol. 2. – P. 8-14. DOI: 10.18653/v1/E17-2002.
 12. Wu Z., Tamkin A., Papadimitriou I. Oolong: Investigating What Makes Transfer Learning Hard with Controlled Studies // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. – Singapore: Association for Computational Linguistics, 2023. – P. 3280-3289. DOI: 10.18653/v1/2023.emnlp-main.198.
 13. Deshpande A., Partha Talukdar P., Narasimhan K. When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Seattle, USA: Association for Computational Linguistics, 2022. – P. 3610-3623. DOI: 10.18653/v1/2022.naacl-main.264.
 14. Lauscher A., Ravishankar V., Vulić I., Glavaš G. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – Association for Computational Linguistics, 2020. – P. 4483-4499. DOI: 10.18653/v1/2020.emnlp-main.363.
 15. Upadhyay S.G., Martinez-Lucas L., Su B.H., Lin W.C., Chien W.S., Wu Y.T., Katz W., Busso C., Lee C.C. Phonetic Anchor-Based Transfer Learning to Facilitate Unsupervised Cross-Lingual Speech Emotion Recognition // ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – Rhodes Island, Greece: IEEE, 2023. DOI: 10.1109/ICASSP49357.2023.10095250.
 16. Feraru S.M., Schuller D., Schuller B. Cross-language acoustic emotion recognition: An overview and some tendencies // 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). – Xi'an, China: IEEE, 2015. – P. 125-131. DOI: 10.1109/ACII.2015.7344561.
 17. Kovacevic L., Bradic V., Melo G., Zdravkovic S., Ryzhova O. Ezglot. – 2022. – URL: <https://ezglot.com>.
 18. Wichmann S., Holman E.W., Bakker D., Brown C.H. Evaluating linguistic distance measures // Physica A: Statistical Mechanics and its Applications. – 2010. – Vol. 389, № 17. – P. 3632-3639. DOI: 10.1016/j.physa.2010.05.011.
 19. Patil V., Talukdar P., Sarawagi S. Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. – Dublin, Ireland: Association for Computational Linguistics, 2022. – Vol. 1. – P. 219-233. DOI: 10.18653/v1/2022.acl-long.18.
 20. Liu C.L., Hsu T.Y., Chuang Y.S., Lee H.Y. A Study of Cross-Lingual Ability and Language-specific Information in Multilingual BERT // arXiv.org. – 2020. DOI: 10.48550/arXiv.2004.09205.
 21. 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. – Vol. 42, № 4. – P. 335-359. DOI: 10.1007/s10579-008-9076-6.
 22. Christop I. nEMO: Dataset of Emotional Speech in Polish // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – Torino, Italia: ELRA and ECCL, 2024. – P. 12111-12116. – URL: <https://aclanthology.org/2024.lrec-main.1059>.
 23. Zhou K., Sisman B., Liu R., Li H. Emotional voice conversion: Theory, databases and ESD // Speech Communication. – 2022. – Vol. 137. – P. 1-18. DOI: 10.1016/j.specom.2021.11.006.
 24. Xin D., Jiang J., Takamichi S., Saito Y., Aizawa A., Saruwatari H. JNVN: A Corpus of Japanese Emotional Speech with Verbal Content and Nonverbal Expressions // IEEE Access. – IEEE, 2024. – Vol. 12. – P. 19752-19764. DOI: 10.1109/ACCESS.2024.3360885.
 25. Boito M.Z., Iyer V., Lagos N., Besacier L., Calapodescu I. mHuBERT-147: A Compact Multilingual HuBERT Model // arXiv.org. – 2024. – P. 3939-3943. DOI: 10.48550/arXiv.2406.06371.
 26. 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. – Vol. 45, № 9. – P. 10745-10759. DOI: 10.1109/TPAMI.2023.3263585.

Таблица 1

Общие сведения о языках

Параметр	Язык				
	русский	английский	польский	китайский	японский
Классификация	Индоевропейская семья – Славянская ветвь – Восточно-славянская группа	Индоевропейская семья – Германская ветвь	Индоевропейская семья – Славянская ветвь – Западно-славянская группа	Сино-тибетская семья	Японо-рюкюская семья
Тип морфологической структуры	Флективный	Аналитический	Флективный	Изолирующий	Агглютинативный
Базовый порядок слов	SVO*	SVO	SVO	SVO	SOV

*SOV и SVO отражают взаимное расположение подлежащего (subject), глагола (verb) и прямого дополнения (object) внутри предложения

Таблица 2

Фонологические признаки (по данным WALS)

Признак	Язык				
	русский	английский	польский	китайский	японский
Инвентарь согласных	Относительно большой	Большой	Большой	Средний	Относительно маленький
Инвентарь гласных	Средний	Средний	Средний	Средний	Средний
Соотношение согласных и гласных	Большое	Маленькое	Относительно большое	Среднее	Среднее
Озвончение в смычных и фрикативных согласных	Есть	Есть	Есть	Только во фрикативных	Есть
Система смычных согласных	/p, t, k, b, d, g/	/p, t, k, b, d, g/	/p, t, k, b, d, g/	/p ^h , t ^h , k ^h , p, t, k /	/p, t, k, b, d, g/
Увулярные согласные	Нет	Нет	Нет	Нет	/N/
Латеральные согласные	/l/	/l/	/l/	/l/	Нет
Велярные носовые	Нет	Нет в начале слова	–	Нет в начале слова	Нет
Назализация гласных	Нет	Нет	–	Нет	Нет
Передние огрубленные гласные	Нет	Нет	Нет	Только высокого ряда	Нет
Структура слога	Сложная	Сложная	Сложная	Умеренно сложная: (C)V(C) как основа	Умеренно сложная: (C)V(C) как основа
Тон	Нет	Нет	Нет	Сложная тоновая система	Простая тоновая система
Фиксированное место ударения	Нет	Нет	Предпоследний слог	Нет	–
Ударение, чувствительное к весу слога	Ударение может быть где угодно	Право-ориентированное: один из трёх последних слогов	Нет (фиксированное ударение)	Непредсказуемое	–
Факторы, влияющие на вес слога	Зависит от лексики	Долгий гласный; наличие согласного на конце слога	Нет	Зависит от лексики	–
Ритмическое ударение	Нет	Хорей	Хорей (каждый чётный слог)	–	–

Признак	Язык				
	русский	английский	польский	китайский	японский
Отсутствие распространённых согласных	Все присутствуют	Все присутствуют	Все присутствуют	Все присутствуют	Все присутствуют
Присутствие редких согласных	Нет	‘th’	Нет	Нет	Нет

Таблица 3

Синтаксические признаки (по данным WALS)

Признак	Язык				
	русский	английский	польский	китайский	японский
Выражение субъекта-местоимения	Обязательное	Обязательное	Клитика на одном из слов	Опциональное	Опциональное
Деление по полу	Три	Три	Три	Нет	Нет
Суффикс множественного числа	Обязателен	Обязателен	Обязателен	Только для людей, опционален	Только для людей, опционален
Грамматические падежи	Семь	Два	Семь	Нет	Семь
Наличие перфекта	Нет	Да	Да	Нет	Нет
Грамматические показатели перфектива/имперфектива	Да	Нет	Да	Нет	Да
Кодирование эвиденциальности	Нет	Нет	Нет	Да	Нет

Приведённые дальше значения основаны на векторах *lang2vec*. Чем ближе значение к нулю, тем выше сходство между языками.

Таблица 4

Синтаксическое различие языков

Язык	Русский	Английский	Польский	Китайский	Японский
Русский	0	0,49	0,47	0,57	0,66
Английский	0,49	0	0,59	0,57	0,66
Польский	0,47	0,59	0	0,6	0,73
Китайский	0,57	0,57	0,6	0	0,59
Японский	0,66	0,66	0,73	0,59	0

Таблица 5

Географическая дистанция между языками

Язык	Русский	Английский	Польский	Китайский	Японский
Русский	0	0,2	0,2	1	0,3
Английский	0,2	0	0,1	1	1
Польский	0,2	0,1	0	1	0,4
Китайский	1	1	1	0	1
Японский	0,3	1	0,4	1	0

Таблица 6

Фонологическое различие языков

Язык	Русский	Английский	Польский	Китайский	Японский
Русский	0	0,28	0	0,62	0,43
Английский	0,28	0	0,28	0,57	0,50
Польский	0	0,28	0	0,62	0,43
Китайский	0,62	0,57	0,62	0	0,62
Японский	0,43	0,50	0,43	0,62	0

Таблица 7

Генетическое различие языков

Язык	Русский	Английский	Польский	Китайский	Японский
Русский	0	0,9	0,4	1	1
Английский	0,9	0	0,9	1	1
Польский	0,4	0,9	0	1	1
Китайский	1	1	1	0	1
Японский	1	1	1	1	0

Таблица 8

Различие в фонетическом инвентаре

Язык	Русский	Английский	Польский	Китайский	Японский
Русский	0	0,56	0,49	0,69	0,54
Английский	0,56	0	0,55	0,6	0,55
Польский	0,49	0,55	0	0,66	0,58
Китайский	0,69	0,6	0,66	0	0,66
Японский	0,54	0,55	0,58	0,66	0

Таблица 9

Различия в признаках между языками (на основе объединения всех предыдущих признаков)

Язык	Русский	Английский	Польский	Китайский	Японский
Русский	0	0,5	0,5	0,6	0,6
Английский	0,5	0	0,5	0,6	0,6
Польский	0,5	0,5	0	0,6	0,6
Китайский	0,6	0,6	0,6	0	0,6
Японский	0,6	0,6	0,6	0,6	0

Материал поступил в редакцию 24.01.25.

Сведения об авторах

ЛЕМАЕВ Владислав Игоревич – аспирант, кафедра теоретической и прикладной лингвистики Московский государственный университет им. М.В. Ломоносова
e-mail: vladzhkv98@mail.ru

ЛУКАШЕВИЧ Наталья Валентиновна – доктор технических наук, профессор, кафедра теоретической и прикладной лингвистики МГУ имени М.В. Ломоносова
e-mail: louk_nat@mail.ru