ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ, АНАЛИЗ ДАННЫХ, МАШИННОЕ ОБУЧЕНИЕ

УДК 004.93'1, 004.934.2, 159.942.33 ISSN 1995-5499
DOI: https://doi.org/
Поступила в редакцию 09.11.2020
Подписана в печать 02.02.2021

ИССЛЕДОВАНИЕ ПОДХОДОВ К КЛАССИФИКАЦИИ ЭМОЦИЙ В НЕВЕРБАЛЬНОМ РЕЧЕВОМ ПОВЕДЕНИИ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

© 2020 М. Ю. Уздяев✉, А. В. Рябинов

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук», Санкт-Петербургский институт информатики и автоматизации Российской академии наук 14 линия В.О., 39, 199178 Санкт-Петербург, Российская Федерация

Аннотация. Распознавание эмоций является актуальной задачей ввиду активного развития систем человеко-машинного взаимодействия и цифровых систем коммуникации. В области автоматического распознавания эмоций исследуется, как правило, поведенческая компонента структуры эмоций, которую проще всего анализировать бесконтактно и без участия испытуемого. Экспрессивная компонента эмоций может быть представлена в различных модальностях: мимические выражения, поза и двигательная активность тела, вербальное и невербальное речевое поведение. Наряду с другими модальностями, невербальное речевое поведение может быть использовано для опосредованного распознавания эмоций. Его анализ становится особенно актуальным в случае недостатка или отсутствия данных других модальностей, а также в моделях многомодального распознавания. В данной статье рассматриваются вопросы распознавания эмоций в речи на основе обработки признаковых представлений записей речи в пространстве признаков eGeMAPS, позволяющем выделить наиболее значимую информацию о невербальном проявлении эмоций в аудиосигнале. Распознавание эмоций выполнялось на следующих наборах данных: CREMA-D, IEMOCAP, Emo-DB, RAVDESS, SAVEE, TESS, а также на их комбинациях. Для предварительной оценки применимости того или иного набора данных в рассматриваемом признаковом пространстве была использована предварительная визуализация данных при помощи алгоритма t-SNE. В качестве методов классификации были выбраны методы, основанные на метрической оценке взаимного расположения данных относительно друг друга: метод *k*-ближайших соседей и метод опорных векторов. В статье приводятся результаты оценки качества классификации исследуемых алгоритмов на основе следующих метрик: доля правильных ответов, точность, полнота. Проведенные эксперименты показали, что метод опорных векторов показывает лучшие результаты в задаче многоклассовой классификации, в то время как метод *k*-ближайших соседей — в задаче бинарной классификации. При распознавании отдельных классов оба метода достигают наибольшую, не ниже 0,55, точность при распознавании «гнева», наименьшую для классов «счастья» и «отвращения».

Ключевые слова: эмоциональные вычисления, распознавание эмоций, визуализация многомерных данных, машина опорных векторов, *k*-ближайших соседей.

✉ Уздяев Михаил Юрьевич
e-mail: m.y.uzdiaev@gmail.com

ВВЕДЕНИЕ

Современные модели, методы и системы распознавания эмоций человека базируются, в основном, на концепции базовых эмоций — выделении некоторого набора элементарных эмоций, не сводимых к другим [1]. Данная концепция принимается в качестве основания для категоризации данных для выполнения дальнейшей классификации системами автоматического распознавания эмоций. Некоторыми исследователями невербальное речевое поведение считается основанием для выделения базовых эмоций [2, 3]. Наборы данных [4, 5], использующие разметку, основанную на концепции базовых эмоций, однако, могут содержать в себе различный набор классов эмоций, использующихся для категоризации данных. Это может быть обусловлено несколькими причинами: применение различных оснований для выделения базовых эмоций, не достаточная представленность той или иной эмоции в наборах, не согласованность при разметке данных экспертами и т. д. Неоднородность наборов данных ставит важнейшую научную задачу сравнительного исследования различных алгоритмов классификации на большом наборе различных наборов данных, содержащих в себе невербальные проявления эмоций человека в речи.

Стоит также особо отметить, что в некоторых специфических задачах на первый план выходят вопросы переразметки известных наборов данных на другие категории. Так, в задаче выявления агрессивного поведения людей, на передний план выходит выявление негативных эмоций, которые обычно сопровождают проявления агрессии [6, 7]. Такая переразметка может быть выполнена простым отнесением экземпляров того или иного класса эмоций к более общей категории — негативные эмоции и остальные эмоции, к которым относят позитивные и нейтральные эмоции. Так, к негативным эмоциям обычно относят гнев, печаль, отвращение, страх, а к остальным — счастье, волнение, скука, удивление, нейтральную эмоцию. Несмотря на всю свою простоту и очевидность, данному подходу уделяется недостаточно внимания в современной научной литературе.

Наряду с обозначенными выше проблемами, стоит также уделить внимание проблеме оценки данных на предмет их группировки в том или ином признаковом пространстве. Такая задача может быть выполнена при помощи различных процедур проекции данных, представленных в признаковом пространстве большой размерности, в признаковое пространство малой размерности с последующей визуализацией данных в этом пространстве. Это позволяет выполнить предварительную качественную оценку данных на предмет их применимости в задаче классификации.

Целью данной работы является выполнение визуализации и качественной оценки данных, содержащих невербальные речевые проявления эмоций для различных наборов данных, а также выполнение сравнительного исследования алгоритмов классификации, обученных на этих наборах. При этом, и визуализация, и классификация выполняются на основании двух способов категоризации данных: использование классов, представленных в наборе данных по умолчанию, а также разбиение на два класса эмоций — негативные и остальные.

1. ОБЗОР ЛИТЕРАТУРЫ

Разработка модели машинного обучения для классификации эмоций по голосу является нетривиальной задачей, поскольку голосовой сигнал содержит в себе много информации, как напрямую относящейся к передаче текущего эмоционального состояния человека, так и вовсе не относящейся к эмоциям. Поэтому решение задачи автоматического распознавания эмоций в первую очередь подразумевает определение некоторого релевантного набора признаков, извлекаемого из записи речи. Признаки, извлекаемые из звукового сигнала, делятся на низкоуровневые дескрипторы (low-level-descriptors, LLD) и функциональные признаки. Низкоуровневые дескрипторы включают в себя просодические (высота тона, громкость, энергия, тембр, про-

должительность пауз и др.) и спектральные (фундаментальная частота, частоты основных формант, мел-кепстральные частотные коэффициенты (MFCC), кепстральные коэффициенты линейного предсказания (LPCC) и др.) характеристики, а также их производные по времени. Функциональные признаки включают в себя статистические показатели низкоуровневых дескрипторов (минимум, максимум, различные процентили, zero-crossing-rate). Рассмотрим некоторые решения в области распознавания эмоций в речи.

Результатом работы авторов [8, 9] стал инструмент OpenSMILE, позволяющий извлекать широкий спектр параметров звукового сигнала, а также применять различные функции к этим параметрам. Благодаря этому появились, например, стандартные наборы признаков для конференций INTERSPEECH, содержащие более 5000 элементов [10], а также попытки унифицировать признаковое пространство для задач аффективных вычислений [11].

Наиболее распространенным подходом в современных исследованиях является извлечение комбинации вышеописанных параметров из акустического сигнала с последующей конкатенацией их в признаковые векторы, которые в дальнейшем используются для обучения классификатора, выбор которого также является важным шагом в решении задачи. К наиболее популярным в задаче распознавания эмоций по голосу классификаторам относятся: машина опорных векторов (Support Vector Machine, SVM), скрытые марковские модели (Hidden Markov Model, HMM), гауссовская смешанная модель (Gaussian Mixture Model, GMM), алгоритм *k*-ближайших соседей (*k*-Nearest Neighbors, *k*-NN), различные архитектуры глубоких нейронных сетей (НС). Так, в работе [12] авторы продемонстрировали высокую степень релевантности просодических характеристик в задаче распознавания эмоций. Их набор признаков состоял из высоты тона, энергии, частот пяти основных формант F1–F5 и был использован для обучения скрытой марковской модели. При этом, точность классификации составила 67,8 %. Данный результат, однако, был получен на собственном наборе данных, содержащем четыре эмоциональные категории. В исследовании [13] были использованы 9 акустических низкоуровневых дескрипторов для обучения глубокой рекуррентной НС с долгой краткосрочной памятью (Long Term Short Memory — LSTM). На наборе данных TUM AVIC для классификации одного из четырех классов получены результаты метрики Unweighted Average Recall 67,6 %, однако данный алгоритм более направлен на выявление сложных аффективных состояний (смех, уверенность, неуверенность, остальное), нежели на выявление базовых эмоций. В работе [14], напротив, извлекали из аудиосигнала исключительно спектральные характеристики: мел-кепстральные частотные коэффициенты, кепстральные коэффициенты линейного предсказания, Perceptual Linear Predictive Cepstrum (PLPC), Mel Frequency Perceptual Linear Predictive Cepstrum (MFPLPC), фундаментальная частота f_0, амплитуда, фаза, а также их статистические функции. Полученный в итоге 64-размерный вектор был использован для обучения глубокой НС. Данная модель позволяет достичь точности распознавания одной из семи эмоций на наборе данных Emo-DB равной 85,7 %. Стоит отметить, что в их работе отсутствует описание использованной архитектуры глубокой НС. Авторами [15] в своем исследовании был использован 39-размерный признаковый вектор на основе MFCC для обучения смешанной гауссовой модели, получив среднюю точность распознавания на наборе данных IEMOCAP 54,34 %. Однако набор данных IEMOCAP не был использован целиком, ограничившись образцами, представленными лишь четырьмя базовыми эмоциями. Наконец, в работе [16] автор извлекал из сигнала просодические и спектральные низкоуровневые дескрипторы, и их статистические функции, получив 43 параметра, из которых с помощью алгоритма RELIEF-F [17] было выбрано 14 наиболее важных. Эти параметры были использованы в итоговом признаковом векторе. Также был проведен сравнительный анализ нескольких алгоритмов классификации: *k*-ближайших соседей, пятислойного

перцептрона и ансамбля нейронных сетей. На собственном наборе данных, содержащем 5 эмоциональных категорий, этими моделями была достигнута средняя взвешенная точность классификации, равная 55 %, 65 % и 70 % соответственно.

Одним из подходов, показывающим лучшие результаты распознавания на данный момент, является построение спектрограмм, их представление в качестве цифровых изображений и использование методов цифровой обработки изображений для решения задачи классификации. Так, авторы работы [18] предложили метод глубокого обучения глубокой сверточной НС на предварительно полученных спектрограммах очищенного от шума звукового сигнала, получив точность классификации, равную 66%, на наборе данных IEMOCAP. Однако, для обучения и тестирования метода авторы использовали не полный набор эмоций, содержащихся в IEMOCAP, ограничившись лишь четырьмя базовыми эмоциями (гнев, счастье, грусть, нейтральная).

Наиболее прогрессивными и эффективными на данный момент являются так называемые end-to-end подходы, которые работают непосредственно с дискретизированным аудиосигналом в формате WAV и в которых предобработка, извлечение релевантных признаков и классификация объединены в единый «черный ящик». Для таких подходов обычно используются глубокие сверточные НС в комбинации с другими архитектурами. В работе [19] была представлена архитектура глубокой НС, состоящей из сверточных и LSTM слоев. Для сравнения эффективности их модели со стандартным эвристическим подходом, был так же извлечен набор признаков, который был использован для обучения SVM и глубокой НС долгой краткосрочной памяти. Предложенный авторами метод превзошел по качеству классификации обе базовые модели, показав результат на наборе данных RECOLA 68 %. Авторы работы [20] описывают метод распознавания эмоций с помощью переноса обучения пятислойной глубокой сверточной НС SoundNet [21]. Авторы рассматривают задачу бинарной классификации «гнев»–«не гнев». В работе использована архитектура SoundNet, у которой веса первых двух сверточных слоев фиксируют, а последующие три инициализируются случайно с целью дообучения рассматриваемой задаче. Получившаяся НС дообучается на наборе данных IEMOCAP. Авторами заявлено улучшение качества классификации в сравнении с обучением с нуля той же НС. Кроме того, отмечается значительное превосходство в генерализации модели, обученной при помощи подхода переноса обучения, над обученной с нуля. Однако в работе отсутствует сравнение предлагаемой модели с другими.

Из рассмотренных выше подходов видно, что проводимые исследования зачастую недостаточно полные, а именно: авторы не проводят тестирование разрабатываемых методов на большом количестве наборов данных, а также не проверяют способность разработанных моделей выделять ту или иную отдельную категорию, представленную в наборах данных. Взаимное расположение данных в выбранном пространстве признаков также часто не учитывается. Исходя из этого нельзя достоверно предположить какие результаты покажет тот или иной подход в случае классификации эмоций, не входящих в базовый набор подхода. В данной работе предпринята попытка сравнить эффективность алгоритмов машинного обучения (k-ближайших соседей и метод опорных векторов) на разных наборах данных, содержащих проявления эмоций в речи, в зависимости от расположения данных в пространстве признаков.

2. ОПИСАНИЕ ПРЕДЛАГАЕМОГО ПОДХОДА

2.1. Выбор признакового пространства

Первой важной задачей в разработке модели машинного обучения для распознавания эмоций по голосу является выбор подходящего признакового пространства. На текущий момент не существует единого мнения относительно релевантности тех или иных параметров акустического сигнала применительно к задаче распознавания эмоциональных состо-

яний по голосу. Как было указано выше, в различных методах применяются различные признаковые пространства, в которых может быть представлен аудиосигнал, содержащий запись человеческого голоса, — чистый аудиосигнал, спектрограммы, основанные на оконном преобразовании Фурье в линейной и мел-частотной шкале, MFCC, LPCC и др. В попытках стандартизации признаковых пространств, используемых в задачах распознавания эмоций по голосу, были разработаны наборы признаков Geneva Minimalistic Acoustic Parameter Set (GeMAPS) и Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [11]. Эти наборы содержат в себе множество параметров звукового сигнала, среди которых: статистические функции от основной частоты и амплитуды, мел-кепстральные коэффициенты (MFCC) 1–4, приближенная оценка количества слов в секунду. Эти параметры наиболее полно отражают основные акустические признаки эмоций [22]. Набор признаков GeMAPS отличается от набора eGeMAPS количеством используемых низкоуровневых дескрипторов. Также особенностью данного набора параметров является независимость размерности итогового признакового вектора от длительности звукового сигнала, что обеспечивает инвариантность признакового пространства относительно продолжительности звучания аудиозаписи, содержащей речь. Для извлечения этих признаков из аудиосигнала был разработан программный продукт OpenSMILE [8].

2.2. Визуализация многомерных данных

Для оценки применимости того или иного набора данных для решения задачи классификации целесообразно сначала оценить взаимное расположение данных различных классов из этого набора в выбранном пространстве признаков. Это может быть выполнено с помощью проекции данных в пространство малой размерности, в котором сохраняется структура расположения экземпляров данных в этом пространстве, с последующей визуализацией данных в новом пространстве. Стоит также отметить, что методы снижения размерности и визуализации данных помимо предварительной качественной оценки данных для решения задачи классификации могут использоваться и в других задачах. Например, в методах переноса обучения для оценки применимости предобученных глубоких нейросетевых моделей, использующихся для выделения признаков, на которых в дальнейшем планируется дообучить НС какой-либо специфичной задаче либо для выбора подходящего признакового представления данных для достижения высоких показателей в различных задачах обучения с учителем — классификации и регрессии. Для снижения размерности был использован алгоритм стохастического вложения соседей с распределением Стьюдента (t-Distributed Stochastic Neighbor Embedding — t-SNE) [23]. Он является одним из самых распространенных методов визуализации многомерных данных. Суть данного метода заключается в проекции данных больших размерностей в пространства, обладающие меньшей размерностью, с сохранением расстояний между экземплярами данных, близко расположенных друг относительно друга. Это позволяет группировать однородные данные в пространствах малой размерности и представлять их в наглядном виде на графиках. Перед применением алгоритма также производится z-нормализация выборки. Полученные таким образом визуальные представления отражают взаимное расположение данных различных классов и позволяют предварительно качественно оценивать обрабатываемые данные на возможность применения их для дальнейшей классификации.

2.3. Алгоритмы классификации

В качестве сравниваемых классификаторов были выбраны: алгоритм k-ближайших соседей и метод опорных векторов. Выбор алгоритмов классификации обусловлен следующими соображениями. Во-первых, выбранные алгоритмы позволяют обрабатывать данные в исходном пространстве признаков без сложных предварительных преобразований данных. Это свойство позволяет оценить

применимость того или иного пространства признаков для классификации эмоций в речи человека. Во-вторых, рассматриваемые алгоритмы позволяют выполнять классификацию данных, представленных в многомерных метрических пространствах признаков на основе анализа взаимных расстояний данных друг относительно друга. Это свойство позволяет соотнести результаты классификации с результатами визуализации данных в пространствах низкой размерности, в которых также оценивается взаимное расположение экземпляров данных на основе их взаимного метрического расположения. В-третьих, данные алгоритмы способны эффективно обрабатывать признаковые пространства, имеющие малый размер векторов признаков, в отличие от других методов, таких как нейронные сети, которые требуют больших репрезентативных наборов данных, векторы признаков которых также должны иметь высокую размерность. Примером такого пространства, векторы признаков которого имеют малый размер, и является пространство eGeMAPS.

Алгоритм *k*-ближайших соседей является алгоритмом непараметрического обучения с учителем. Суть работы данного алгоритма заключается в следующем: алгоритм определяет класс объекта z_i путем определения класса большинства объектов z_j из числа *k*-ближайших соседей объекта z_i. Для оценки расстояния между объектами в алгоритме *k*-ближайших соседей могут использоваться различные меры расстояния, такие как евклидова мера, манхэттенская мера, косинусная мера и др [24].

Метод опорных векторов является линейным методом классификации. Основная идея метода опорных векторов, заключается в построении разделяющей гиперплоскости между данными в случае бинарной классификации и разделяющих гиперплоскостей в случае многоклассовой классификации. Наличие таких гиперплоскостей говорит о линейной разделимости данных в признаковом пространстве [25, 26].

Выбранные алгоритмы классификации основаны на оценке евклидового расстояния в признаковом пространстве, поэтому масштабирование данных влияет на качество классификации как для *k*-NN [27], так и для SVM [28]. В связи с этим, с целью приведения данных к единому масштабу значений все векторы признаков были подвергнуты z-нормализации. Этот метод представляет из себя масштабирование данных на основе среднего по выборке значения и стандартного отклонения.

3. ОПИСАНИЕ ЭКСПЕРИМЕНТОВ

3.1. Описание наборов данных

Для исследования возможности классификации эмоций в аудиосигнале были выбраны следующие наборы данных: IEMOCAP (Interactive emotional dyadic motion capture database) [29], CREMA-D (Crowd-sourced emotional multimodal actors dataset) [30], Emo-DB [5], RAVDESS [4], SAVEE [31], TESS [32]. Основные характеристики рассматриваемых наборов данных, а именно название, язык, на котором произносились фразы, половозрастной состав дикторов (количество различных дикторов, их пол и возраст), тип произносимых фраз, количество записей. Технические данные аудиосигнала (формат аудиофайлов, частота дискретизации, количество каналов), представленные в наборе данных классы эмоций сведены в табл. 1. При этом, в рассматриваемых наборах данных представлены следующие классы эмоций: ANG — Anger (гнев); HAP – Happiness (радость); SAD — Sadness (печаль); NEU — Neutral (нейтральная эмоция); DIS — Disgust (отвращение); FEA — Fear (страх); BOR — Boredom (скука); SUR — Surprise (удивление); EXC — Excitement (возбуждение); FRU — Frustration (негодование); CAL — Calm (спокойствие).

Задача распознавания эмоций по голосу осложнена отсутствием единой методики разметки наборов данных. В этом случае для исследования применимости для задачи классификации того или иного пространства признаков, в котором могут быть представлены эмоции в речи, на различных наборах данных целесообразно выполнить следующие действия.

1) Для каждого набора провести переразметку данных, выделяя два класса эмоций — негативные (Negative — NEG) и остальные (REST). При этом, к негативным эмоциям можно отнести эмоции ANG, SAD, FRU, DIS, FEA, а к остальным — эмоции HAP, EXC, NEU, BOR, SUR, CAL.

2) Путем слияния всех наборов данных на английском языке, получить обобщенный набор данных English Assembly, содержащий 19462 образца и произвести попытку как мультиклассовой, так и бинарной классификации. При этом, для мультиклассовой классификации принято решение ограничиться образцами, помеченными шестью самыми распространенными базовыми эмоциями согласно Ekman [32]: ANG, HAP, DIS, FEA, NEU, SAD. В случае бинарной классификации используются все образцы.

3) Выполнить снижение размерности пространства признаков для каждого набора данных, переразмеченных бинарной разметкой наборов и на объединенного набора English Assembly при помощи t-SNE для визуализации взаимного расположения данных и предварительной оценки применимости рассматриваемых наборов для классификации эмоций в речи.

4) Выполнить обучение алгоритмов *k*-NN и SVM задаче классификации на каждом отдельном наборе данных, на переразмеченных бинарной разметкой наборах и на объединенном наборе English Assembly с последующим сравнением результатов классификации и сопоставлением их с визуализациями, полученными при помощи t-SNE.

Таблица 1. Наборы данных, подлежащие исследованию
[Table 1. Datasets to be considered]

Название	Язык	Данные о дикторах	Тип фраз	Количество записей	Формат аудио	Базовая дискретная разметка
CREMA-D	Англ.	48 мужчин и 43 женщины, от 20 до 74 лет, средний возраст — 36 лет	12 коротких предложений	7442	wav, 48 кГц, 2 канала	ANG, HAP, SAD, NEU, DIS, FEA
IEMOCAP	Англ.	5 мужчин и 5 женщин	Диалоги по сценарию и спонтанные импровизации	7304	wav, 48 кГц, 2 канала	ANG, HAP, SAD, NEU, EXC, FRU
Emo-DB	Нем.	5 мужчин и 5 женщин	5 коротких и 5 длинных предложений	535	wav, 16 кГц, 1 канал	ANG, HAP, SAD, NEU, DIS, FEA, BOR
RAVDESS	Англ.	12 мужчин и 12 женщин, от 21 до 33 лет, средний возраст — 26 лет	2 коротких предложения	1440	wav, 48 кГц, 2 канала	ANG, HAP, SAD, NEU, DIS, FEA, SUR, CAL
SAVEE	Англ.	4 мужчины, от 27 до 31 года	120 различных предложений	480	wav, 44,1 кГц, 2 канала	ANG, HAP, SAD, NEU, DIS, FEA, SUR
TESS	Англ.	2 женщины, 26 и 64 года	‘Say the word X’, где X — одно из 200 односложных слов	2800	wav, 48 кГц, 2 канала	ANG, HAP, SAD, NEU, DIS, FEA, SUR

3.2. Результаты визуализации распределения данных в пространстве eGeMAPS

В данной работе была использована реализация алгоритма t-SNE, представленная в библиотеке ScikitLearn [34] языка программирования Python. Чтобы не перегружать визуализацию и не усложнять работу алгоритма, количество подаваемых на вход алгоритма образцов данных было сокращено до случайной выборки, состоящей из 2000 (для объединенного набора данных English Assembly — из 5000) образцов с сохранением пропорций распределения классов эмоций. Визуализации результатов t-SNE на исходных наборах данных при оригинальной разметке представлены на рис. 1.

Исходя из полученных распределений данных в пространстве малой размерности можно сделать следующие выводы. В наборах данных CREMA-D (рис. 1, а), IEMOCAP (рис. 1, в), RAVDESS (рис. 1, г) наблюдаются следующие явления: по большей части, для этих наборов формируется одна единая группа векторных представлений. При этом, экземпляры некоторых классов (ANG, SAD) имеют большую плотность распределения в определенных зонах этой группы, а такие классы, как FEA, NEU распределены по всей области более равномерно. Набор данных SAVEE (рис. 1, д) формирует две группы данных. В обеих группах локализуются экземпляры нескольких классов, однако внутри каждой из этих групп можно выделить области, где наблюдается наиболее высокая концентрация экземпляров одного из классов. В наборе данных Emo-DB (рис. 1, б) данные группируются в соответствии с классами, однако четких границ между этими группами нет, а различные классы распределены с различной плотностью. В наборе TESS (рис. 1, е) наблюдается наиболее ярко выраженная структура среди всех полученных визуализаций, данные разбиты на плотные скопления с четкими границами, причем каждая эмоциональная категория представлена двумя группами. Можно утверждать, что данные сгруппированы не

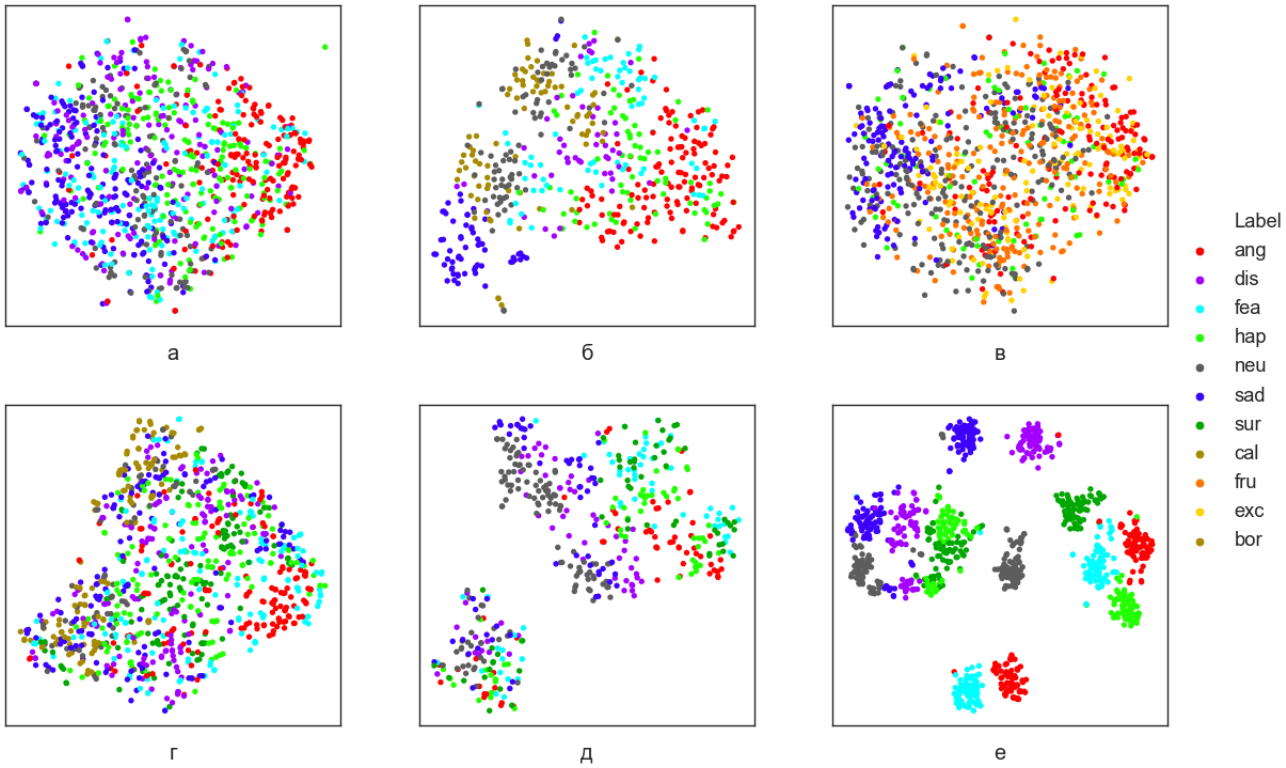

Рис. 1. Визуализации результатов t-SNE на наборах данных при оригинальной разметке: а — CREMA-D, б — Emo-DB, в — IEMOCAP, г — RAVDESS, д — SAVEE, е — TESS
[Fig. 1. Visualization of t-SNE results on datasets by original labeling: а — CREMA-D, б — Emo-DB, в — IEMOCAP, г — RAVDESS, д — SAVEE, е — TESS]

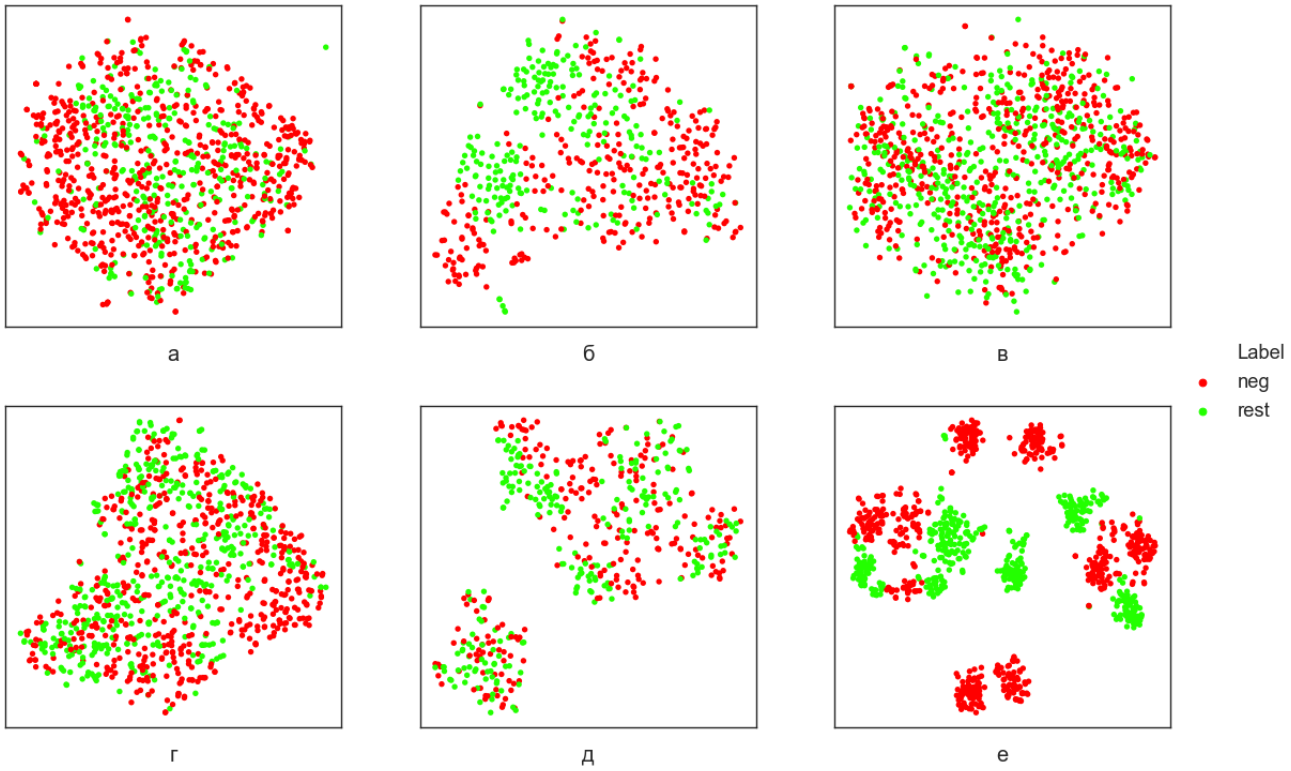

Рис. 2. Визуализации результатов t-SNE на наборах данных при бинарной разметке: а — CREMA-D, б — Emo-DB, в — IEMOCAP, г — RAVDESS, д — SAVEE, е — TESS
[Fig. 2. Visualization of t-SNE results on datasets by binary labeling: а — CREMA-D, б — Emo-DB, в — IEMOCAP, г — RAVDESS, д — SAVEE, е — TESS]

только по представленным классам базовых эмоций, но и по актерам, участвовавшим в записи. Визуализации результатов t-SNE на исходных наборах данных при бинарной разметке представлены на рис. 2.

В наборах данных IEMOCAP (рис. 2, в) и SAVEE (рис. 2, д) экземпляры обоих классов распределены равномерно по всему пространству, нет яркой выраженности скоплений экземпляров данных одного класса по сравнению с другим в той или иной области сформированного при помощи алгоритма t-SNE пространства признаков малой размерности. При визуализации наборов данных RAVDESS (рис. 2, г) и CREMA-D (рис. 2, а) наблюдаются области пространства с преобладанием представленности экземпляров одного класса над экземплярами другого, однако ни линейной разделимости, ни каких бы то ни было других четких границ между классами не наблюдается. В наборах TESS (рис. 2, е) и Emo-DB (рис. 2, б) явно наблюдаются области пространства, которые образуют группы скопления экземпляров данных. Однако, сформированные группы также нельзя разграничить линейно. Визуализации результатов t-SNE на объединенном наборе данных English Assembly при многоклассовой и бинарной разметке набора данных представлены на рис. 3.

На рисунках наглядно продемонстрировано, что как в случае многоклассовой разметки (рис. 3, а), так и в случае бинарной разметки (рис. 3, б) данные сгруппированы, по большей части, в пространстве не по классам. В результате визуализации, отражающей взаимное расположение экземпляров данных друг относительно друга в пространстве признаков eGeMAPS, можно предположить, что алгоритмы классификации, обрабатывающие данные в этом признаковом пространстве, будут иметь относительно низкие результаты вследствие того, что данные различных классов либо группируются в одну большую группу, либо происходит группировка данных по признакам, отличным от принадлежности данных тому или иному классу, кроме набора TESS.

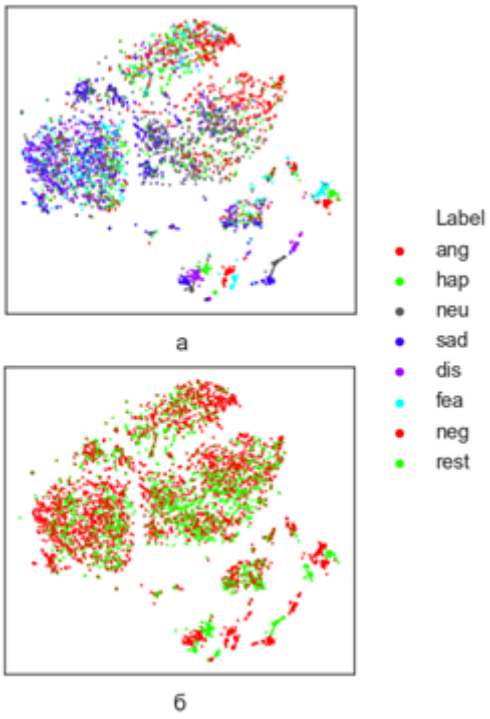

Рис. 3. Визуализация результатов t-SNE на объединенном наборе данных English Assembly: а — многоклассовая разметка, б — бинарная разметка
[Fig. 3. Visualization of t-SNE results on the compound dataset English Assembly: а — multiclass labeling, б — binary labeling]

3.3. Классификация

Для обучения алгоритмов *k*-NN и SVM данные в каждом наборе были предварительно разбиты на обучающую и тестовую выборку в пропорции 7:3, с сохранением пропорций распределения классов в выборках. В качестве метрик качества классификации выбраны следующие: доля верно распознанных экземпляров $acc = (TP + TN) / (TP + TN + FP + FN)$, мера точности $pr = TP/(TP + FP)$ и мера полноты $rec = TP / (TP + FN)$, где *TP* (True Positive), *FP* (False Positive), *FN* (False Negative) — соответственно количество истинно-положительных, ложноположительных и ложноотрицательных экземпляров данных, распознанных тестируемой моделью. Таким образом, интуитивно, *rec* демонстрирует способность алгоритма обнаруживать определенный класс, а *pr* — способность отличать этот класс от других классов. Обучение и тестирование алгоритмов классификации *k*-NN и SVM также выполнялась с использованием библиотеки ScikitLearn языка программирования Python. В экспериментах использовался алгоритм SMV с линейным ядром. В табл. 2–3 представлены результаты распознавания эмоций для наборов данных CREMA-D, Emo-DB, IEMOCAP, RAVDESS, SAVEE, English Assembly и Tess.

Из приведенных данных видно, что на многоклассовой классификации алгоритм *k*-NN для наборов CREMA-D, IEMOCAP, SAVEE, English Assembly лучше всего распознает метку гнева, для Emo-DB — метку грусти, для RAVDESS — метку удивления и спокойствия. Хуже всего распознает для SAVE, CREMA-D — отвращение, RAVDESS и Emo-DB — счастье, IEMOCAP — счастье, не сумев правильно распознать ни один экземпляр этого класса, а English Assembly — эмоции отвращения и счастья. На бинарной же классификации алгоритм *k*-NN для наборов CREMA-D, English Assembly показывает лучшие результаты при распознавании класса «отрицательные эмоции», в то время как для наборов Emo-DB, IEMOCAP, RAVDESS, SAVEE алгоритм достиг более высокой точности (*pr*) при распознавании класса «остальные эмоции», а более высокой полноты (*rec*) при распознавании класса «отрицательные эмоции».

Классификатор SVM на многоклассовой классификации показал лучшие результаты при распознавании эмоции скуки для набора Emo-DB, гнева для CREMA-D, IEMOCAP, SAVEE, English Assembly, меток удивления и гнева для RAVDESS. Худшие — при распознавании счастья для наборов CREMA-D, Emo-DB, IEMOCAP, грусти для RAVDESS и SAVEE, эмоции отвращения для English Assembly. На бинарной же классификации алгоритм SVM имеет чуть более низкие показатели, чем *k*-NN для всех наборов и также достигает большей точности (*pr*) при распознавании класса «остальные эмоции», а большей полноты (*rec*) при распознавании класса «отрицательные эмоции» для наборов данных Emo-DB, IEMOCAP, SAVEE, а для CREMA-D, RAVDESS, English Assembly лучше справляет-

Таблица 2. Результаты распознавания для наборов данных CREMA-D, Emo-DB, IEMOCAP, RAVDESS
[Table 2. Recognition results for datasets CREMA-D, Emo-DB, IEMOCAP, RAVDESS]

Класс	CREMA-D				Emo-DB				IEMOCAP				RAVDESS			
	k-NN		SVM		*k*-NN		SVM		*k*-NN		SVM		*k*-NN		SVM	
	pr	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*
	многоклассовая															
ang	0,60	0,73	0,70	0,74	0,60	0,84	0,68	0,59	0,55	0,43	0,60	0,43	0,56	0,73	0,77	0,71
cal													0,61	0,90	0,76	0,77
bor					0,77	0,50	0,94	0,75								
exc									0,37	0,15	0,46	0,37				
fru									0,37	0,59	0,41	0,57				
dis	0,41	0,21	0,45	0,42	0,44	0,36	0,91	0,91					0,55	0,63	0,62	0,65
fea	0,47	0,26	0,52	0,43	0,78	0,41	0,78	0,82					0,63	0,54	0,60	0,69
hap	0,52	0,37	0,50	0,47	0,43	0,17	0,43	0,50	0,00	0,00	0,26	0,05	0,37	0,33	0,57	0,67
neu	0,38	0,69	0,51	0,58	0,54	1,00	0,78	0,90	0,51	0,43	0,47	0,52	0,43	0,38	0,62	0,33
sad	0,49	0,64	0,55	0,63	1,00	0,88	0,88	0,94	0,51	0,78	0,59	0,64	0,64	0,29	0,54	0,46
sur													0,70	0,67	0,77	0,83
μ	0,48	0,48	0,54	0,55	0,65	0,59	0,77	0,77	0,38	0,40	0,46	0,43	0,56	0,56	0,66	0,64
acc	0,48		0,54		0,63		0,75		0,45		0,48		0,57		0,66	
	бинарная															
neg	0,76	0,92	0,75	0,90	0,80	0,93	0,82	0,86	0,64	0,81	0,69	0,75	0,76	0,84	0,74	0,73
rest	0,69	0,36	0,63	0,36	0,89	0,69	0,80	0,76	0,66	0,45	0,66	0,60	0,79	0,69	0,69	0,70
μ	0,72	0,64	0,69	0,63	0,84	0,81	0,81	0,81	0,65	0,63	0,68	0,67	0,56	0,56	0,66	0,64
acc	0,75		0,73		0,83		0,81		0,65		0,68		0,77		0,72	

Таблица 3. Результаты распознавания для наборов данных SAVEE, English Assembly и TESS
[Table 3. Recognition results for datasets SAVEE, English Assembly and TESS]

Класс	SAVEE				English Assembly				TESS			
	k-NN		SVM		*k*-NN		SVM		*k*-NN		SVM	
	pr	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*
	многоклассовая											
ang	0,67	0,80	0,75	0,80	0,69	0,75	0,71	0,73	0,95	0,99	0,98	1,00
dis	0,54	0,47	0,61	0,73	0,55	0,42	0,55	0,54	0,97	0,95	0,98	0,98
fea	0,63	0,67	0,86	0,40	0,57	0,45	0,57	0,51	1,00	0,97	1,00	0,99
hap	0,69	0,60	0,73	0,53	0,59	0,38	0,55	0,37	0,91	0,97	0,99	0,95
neu	0,59	0,77	0,78	0,83	0,55	0,71	0,60	0,72	1,00	1,00	1,00	1,00
sad	0,67	0,40	0,53	0,53	0,58	0,67	0,62	0,68	0,98	0,98	0,99	0,98
sur	0,58	0,47	0,67	0,93					0,96	0,90	0,95	0,99
μ	0,56	0,56	0,66	0,64	0,56	0,56	0,66	0,64	0,97	0,97	0,98	0,98
acc	0,62		0,70		0,59		0,61		0,97		0,98	
	бинарная											
neg	0,57	0,78	0,63	0,77	0,74	0,84	0,69	0,82	0,99	0,99	0,97	0,98
rest	0,66	0,42	0,70	0,55	0,70	0,55	0,62	0,45	0,99	0,99	0,97	0,96
μ	0,56	0,56	0,66	0,64	0,56	0,56	0,66	0,64	0,99	0,99	0,97	0,97
acc	0,60		0,66		0,73		0,67		0,99		0,97	

ся с распознаванием класса «отрицательные эмоции», чем класса «остальные эмоции».

Для набора данных TESS оба классификатора показали близкие к стопроцентным результаты доли верно распознанных экземпляров всех классов как для многоклассовой, так и для бинарной классификции. Это объясняется особенностями данного набора: отсутствием разнообразия дикторов и произносимых фраз (в записи принимали участие 2 женщины, произносившие одну и ту же фразу), а также сильно выраженной искусственной артикуляцией эмоций, одинаковой в пределах одного класса. Полученные результаты классификации соотносятся с t-SNE визуализацией распределения объектов в этом наборе данных, в которой видно четкое разделение на группы по классам.

4. ИНТЕРПРЕТАЦИЯ И ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Обобщая результаты классификации, можно сделать следующие выводы: подтвердилось предварительное предположение, основанное на визуализациях t-SNE, о том, что результаты алгоритмов *k*-NN и SVM покажут относительно низкие результаты на всех наборах, где не наблюдалась группировка данных по классам. На наборах данных, для которых визуализации t-SNE показали лучшее разделение и группировку данных по классам, получены лучшие результаты классификации, к которым можно отнести Emo-DB и TESS. Действительно, лучшие результаты многоклассовой и бинарной классификации оба алгоритма классификации показали на этом наборе данных, что объясняется его особенностями: малым количеством дикторов, крайне ограниченным набором произносимых фраз, ярко выраженной артикуляцией, с которой произносятся фразы. Худшие результаты многоклассовой классификации как *k*-NN, так и SVM продемонстрировали на наборе данных IEMOCAP. Это так же можно объяснить особенностями этого набора данных, а именно: большим разнообразием дикторов; наличием образцов как искусственно смоделированных, так и натуральных эмоциональных состояний дикторов, записанных в импровизационных диалогах. Худшие результаты бинарной классификации оба алгоритма продемонстрировали на наборе данных SAVEE. Кроме того, на некоторых наборах данных (SAVEE, TESS) t-SNE группирует данные по отличному от эмоциональной окраски признаку. Это могут быть кластеры для говорящих разного пола, или для разных актеров. Применение дополнительного разделения по полам или использование идентификации говорящего может улучшить качество классификации.

В целом, можно отметить, что классификатор SVM показывает лучшие, чем классификатор *k*-NN, результаты многоклассовой классификации, и сравнимые с *k*-NN результаты бинарной классификации. Однако, алгоритм *k*-NN показывает лучшие результаты на наборах данных RAVDESS, CREMA-D, EmoDB, TESS и English Assembly в задаче бинарной классификации. Рассматривая способности алгоритмов *k*-NN и SVM распознавать отдельные классы, входящие в исследуемые наборы данных, можно сделать вывод, что в большинстве случаев, класс «гнев» имел наибольшую точность (precision) распознавания среди различных наборов данных, никогда не опускаясь ниже 0,55. Хуже других на большинстве наборов данных распознаются классы счастья и отвращения.

ЗАКЛЮЧЕНИЕ

Снижение размерности данных при помощи t-SNE и их последующая визуализация в пространстве малой размерности позволяют оценить структуру взаимного расположения данных различных классов. Это, в свою очередь, позволяет предварительно оценить результаты классификации различных алгоритмов, обученных на этих данных. Действительно, свои лучшие результаты и алгоритм *k*-NN, и алгоритм SVM показали на наборе данных TESS, у которого наблюдается наиболее четкое разделение данных по классам на визуализации, а худшие — на наборах IEMOCAP и SAVEE, у которых не наблюдается какой-либо структуры на представленных визуализациях.

Кроме того, и результаты визуализации, и результаты классификации, во многом, зависят от особенностей обучающего набора данных. Стоит также добавить, что при многоклассовой классификации на всех наборах данных алгоритм SVM показывает более высокие результаты, чем алгоритм *k*-NN, однако этот алгоритм показывает зачастую более высокие результаты в задаче бинарной классификации. В дальнейшей работе планируется исследовать другие признаковые пространства (спектрограммы аудиосигнала речи человека или MFCC), а также другие методы классификации (байесовские классификаторы, сверточные и реккурентные НС, ансамбли различных алгоритмов классификации, применение методов снижения размерности и т.д.).

БЛАГОДАРНОСТИ

Работа выполнена при поддержке РФФИ (18-29-22061_мк).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Turner, W.* What's basic about basic emotions? / W. Turner, A. Ortony // Psychological review. – 1990. – Vol. 97. – No. 3. – P. 315–331.

2. *Scherer, K. R.* Vocal affect expression: A review and a model for future research / K. R. Scherer // Psychological bulletin. – 1986. – Vol. 99. – No. 2. – P. 143.

3. *Banse, R.* Acoustic profiles in vocal emotion expression / R. Banse, K. R. Scherer // Journal of personality and social psychology. – 1996. – Vol. 70. – No. 3. – P. 614.

4. *Livingstone, S. R.* The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English / S. R. Livingstone, F. A. Russo // PloS one. – 2018. – Vol. 13. – No. 5. – P. e0196391.

5. *Burkhardt, F.* A database of German emotional speech / F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss // In 9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech). – Lisbon, Portugal, 2005. – No. 4. – P. 1517–1520.

6. *Комалова, Л. Р.* Перцептивно-слуховой профиль (образ) агрессора / Л. Р. Комалова // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2016. – No.7 (746). – P. 116–126.

7. *Комалова, Л. Р.* Сопоставление слухового и зрительного видов восприятия агрессивного речевого поведения / Л. Р. Комалова // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2016. – No. 15 (754). – P. 114–128.

8. *Eyben, F.* Opensmile: the munich versatile and fast open-source audio feature extractor / F. Eyben, M. Wöllmer, B. Schuller // Proceedings of the 18th ACM international conference on Multimedia. – 2010. – P. 1459–1462.

9. *Eyben, F.* Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor / F. Eyben, F. Weninger, F. Gross, B. Schuller // Proceedings of the 21st ACM international conference on Multimedia. – 2013. – P. 835–838.

10. *Schuller, B.* The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism / B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, S. Kim // Proceedings of the Annual Conference of the International Speech Communication Association. – INTERSPEECH, 2013. – P. 148–152.

11. *Eyben, F.* The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing / S. Frühholz, J. Dietziker, M. Staib, W. Trost // In IEEE Transactions on Affective Computing. – 2016. – Vol. 7, No. 2. – P. 190–202.

12. *Lin, J. C.* Semi-coupled hidden Markov model with state-based alignment strategy for audio-visual emotion recognition / J. C. Lin, C. H. Wu, W. L. Wei // In Proc. Affective Com-

puting and Intelligent Interaction (ACII). – 2011. – P. 185–194.

13. *Eyben, F.* Audiovisual vocal outburst classification in noisy acoustic conditions / F. Eyben, S. Petridis, B. Schuller, M. Pantic // in ICASSP. – 2012. – P. 5097–5100.

14. *Lalitha, S.* Emotion detection using perceptual based speech features / S. Lalitha, S. Tripathi // 2016 IEEE Annual India Conference (INDICON). – IEEE, 2016. – P. 1–5.

15. *Metallinou, A.* Audio-visual emotion recognition using Gaussian mixture models for face and voice / A. Metallinou, S. Lee, S. Narayanan // In Proc. Int. Symp. Multimedia. – 2008. – P. 250–257.

16. *Petrushin, V. A.* Emotion recognition in speech signal: experimental study, development, and application / V. A. Petrushin // In: Proceedings of ICSLP. – 2000. – P. 222–225.

17. *Kononenko, I.* Estimating attributes: Analysis and extension of RELIEF / I. Kononenko // European conference on machine learning. – Springer, Berlin, Heidelberg, 1994. – P. 171–182.

18. *Satt, A.* Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms / A. Satt, S. Rozenberg, R. Hoory // Interspeech. – 2017. – P. 1089–1093.

19. *Trigeorgis, G.* Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Recurrent Network / G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, S. Zafeiriou // 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2016. – P. 5200–5204.

20. *Elshaer, M. E.* A. Transfer learning from sound representations for anger detection in speech / M. E. A. Elshaer, S. Wisdom, T. Mishra // arXiv preprint arXiv:1902.02120. – 2019.

21. *Aytar, Y.* Soundnet: Learning sound representations from unlabeled video / Y. Aytar, C. Vondrick, A. Torralba // Advances in neural information processing systems. – 2016. – P. 892–900.

22. *Zeng, Z.* A survey of affect recognition methods: audio, visual, and spontaneous expressions / Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang // IEEE Trans. Pattern Anal. Mach. Intell. – 2009. – No. 31(1). – P. 39–58.

23. *Maaten, L.* Visualizing data using t-SNE / L. Maaten, G. Hinton // Journal of machine learning research. – 2008. – Vol. 9, No. Nov. – P. 2579–2605.

24. *Chomboon, K.* An empirical study of distance metrics for k-nearest neighbor algorithm / K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, N. Kerdprasop // Proceedings of the 3rd international conference on industrial application engineering. – 2015. – P. 280–285.

25. *Вапник, В. Н.* Теория распознавания образов / В. Н. Вапник, А. Я. Червоненкис. – Москва: Наука, 1974. – 416 с.

26. *Vapnik, V.* The Nature of Statistical Learning Theory / V. Vapnik – Springer Science & Business Media – 1999, 314 p.

27. *Peterson, L. E.* K-nearest neighbor / L. E. Peterson / Scholarpedia. – 2009. – Vol. 4, No. (2). – P. 1883.

28. *Tax, D. M.* Feature scaling in support vector data descriptions / D. M. Tax, R. P. Duin // Learning from Imbalanced Datasets. – 2000. – P. 25–30.

29. *Busso, C.* IEMOCAP: Interactive emotional dyadic motion capture database / C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, S. Narayanan // Language resources and evaluation. – 2008. – Vol. 42. – No. 4. – P. 335.

30. *Cao, H.* CREMA-D: Crowd-sourced emotional multimodal actors dataset / H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, R. Verma // IEEE transactions on affective computing. – 2014. – Vol. 5, No. 4. – P. 377–390.

31. *Jackson, P.* Surrey audio-visual expressed emotion (savee) database / P. Jackson, S. Haq. – University of Surrey: Guildford, UK – 2014.

32. *Pichora-Fuller, M. K.* Toronto emotional speech set (TESS) / M. K. Pichora-Fuller, K. Dupuis // Scholars Portal Dataverse. – 2020.

33. *Ekman, P.* Basic emotions / P. Ekman. – In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion. Chichester: Wiley. – 1999.

34. *Pedregosa, F.* Scikit-learn: Machine learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas // The Journal of machine Learning research. – 2011. – Vol. 12. – P. 2825–2830.

Уздяев Михаил Юрьевич — младший научный сотрудник лаборатории технологий больших данных социокиберфизических систем, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук.
E-mail: m.y.uzdiaev@gmail.com
ORCID iD: https://orcid.org/0000-0002-7032-0291

Артем Валерьевич Рябинов — программист лаборатории автономных робототехнических систем, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук.
E-mail: iamryabinov@gmail.com
ORCID iD: https://orcid.org/0000-0002-3572-4493

DOI: https://doi.org/
Received 09.11.2020
Accepted 02.02.2021

ISSN 1995-5499

ANALYSIS OF THE APPROACHES TO CLASSIFYING EMOTIONS IN NON-VERBAL COMMUNICATIONS BASED ON MACHINE LEARNING

© 2020 M. Yu. Uzdiaev✉, A. V. Ryabinov

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS),
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
39, 14th Line, 199178 St. Petersburg, Russian Federation

Annotation. Due to the active development of human-machine interaction systems and digital communication channels, emotion recognition is a very important problem. Researchers working on automated emotion recognition usually focus on the behavioural component of emotions, since it can be analysed remotely, not involving the subject. The expressive component can be represented by various modalities: facial expressions, posture and body movements, verbal and non-verbal behaviour. Non-verbal behaviour, alongside with other modalities, can be used for the indirect recognition of emotions. Analysis of this modality becomes particularly relevant, when there is little or no data from the other modalities, as well in multimodal recognition models. The article considers an approach to emotion recognition during communications, based on the processing of feature representations of speech recordings in the eGeMAPS feature set, which allows to determine the most relevant information about non-verbal emotion expression in an audio signal. Emotion recognition was performed using the following datasets: CREMA-D, IEMOCAP, Emo-DB, RAVDESS, SAVEE, and TESS, as well as their combinations. For the preliminary assessment of applicability of a certain data set in the feature space considered, preliminary data visualisation with t-SNE algorithm was used. For classification purposes the methods were selected based on the metric assessment of mutual data distribution: the method of k-nearest neighbours and support vector machines method. The article presents the results of classification of the analysed algorithms, based on the following metrics: percentage of correct answers, accuracy, and completeness. The conducted experiments demonstrated that the support vector machines method performs better for multiclass classification, whereas the k-nearest neighbour method is better for binary classification. When recognising individual classes both methods yield

✉ Uzdiaev Mikhail Yu.
e-mail: m.y.uzdiaev@gmail.com

the maximum accuracy (0.55 or higher) for “anger”; the minimum accuracy was observed for “happiness” and “disgust”.
Keywords: computation of emotions, emotion recognition, visualisation of multidimensional data, support vector machine, k-nearest neighbours.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Turne, W., Ortony A.* What’s basic about basic emotions? // Psychological review. 1990. Vol. 97. No. 3. P. 315–331.

2. *Scherer K. R., Johnstone T., Klasmeyer G.* Vocal expression of emotion // Oxford University Press, 2003. P. 433–456.

3. *Banse R., Scherer K. R.* Acoustic profiles in vocal emotion expression // Journal of personality and social psychology. 1996. Vol. 70. No. 3. P. 614.

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. Vol. 13. No. 5. P. e0196391.

5. *Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B.* A database of German emotional speech // In 9th European Conference on Speech Communication and Technology (Interspeech’2005 — Eurospeech). Lisbon, Portugal, 2005. No. 4. P. 1517–1520.

6. *Komalova L. P.* Auditory-perceptual profile (image) of an aggressor // Bulletin of the Moscow State Linguistic University. Humanitarian sciences. 2016. No. 7 (746). P. 116–126.

7. *Komalova L. P.* Comparing auditory and visual typesof perception of aggressive verbal behavior // Bulletin of the Moscow State Linguistic University. Humanitarian sciences. 2016. No. 15 (754). P. 114–128.

8. *Eyben F., Wöllmer M., Schuller B.* Opensmile: the munich versatile and fast open-source audio feature extractor // Proceedings of the 18th ACM international conference on Multimedia. 2010. P. 1459–1462.

9. *Eyben F., Weninger F., Gross F., Schuller B.* Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor // Proceedings of the 21st ACM international conference on Multimedia. 2013. P. 835–838.

10. *Schuller B., Steidl S., Batliner A., Vinciarelli A., Scherer K., Ringeval F., Chetouani M., Weninger F., Eyben F., Marchi E., Mortillaro M., Salamin H., Polychroniou A., Valente F., Kim S.* The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism // Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH, 2013. P. 148–152.

11. *Eyben F., Dietziker J., Staib M., Trost W.* The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing // In IEEE Transactions on Affective Computing. 2016. Vol. 7. No. 2. P. 190–202.

12. *Lin J. C., Wu C. H., Wei W. L.* Semi-coupled hidden Markov model with state-based alignment strategy for audio-visual emotion recognition // In Proc. Affective Computing and Intelligent Interaction (ACII). 2011. P. 185–194.

13. *Eyben F., Petridis S., Schuller B., Pantic M.* Audiovisual vocal outburst classification in noisy acoustic conditions // in ICASSP. 2012. P. 5097–5100.

14. *Lalitha S., Tripathi S.* Emotion detection using perceptual based speech features // 2016 IEEE Annual India Conference (INDICON). IEEE, 2016. P. 1–5.

15. *Metallinou A., Lee S., Narayanan S.* Audio-visual emotion recognition using Gaussian mixture models for face and voice // In Proc. Int. Symp. Multimedia. 2008. P. 250–257.

16. *Petrushin V. A.* Emotion recognition in speech signal: experimental study, development, and application // In: Proceedings of ICSLP. 2000. P. 222–225.

17. *Kononenko I.* Estimating attributes: Analysis and extension of RELIEF // European conference on machine learning. Springer, Berlin, Heidelberg, 1994. P. 171–182.

18. *Satt A., Rozenberg S., Hoory R.* Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms // Interspeech. 2017. P. 1089–1093.

19. *Trigeorgis G., Ringeval F., Brueckner R., Marchi E., Nicolaou M., Schuller B., Zafeiriou S.* Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Recurrent Network // 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016. P. 5200–5204.

20. *Elshaer M. E. A., Wisdom S., Mishra T.* Transfer learning from sound representations for anger detection in speech // arXiv preprint arXiv:1902.02120. 2019.

21. *Aytar Y., Vondrick C., Torralba A.* Soundnet: Learning sound representations from unlabeled video // Advances in neural information processing systems. 2016. P. 892–900.

22. *Zeng Z., Pantic M., Roisman G. I., Huang T. S.* A survey of affect recognition methods: audio, visual, and spontaneous expressions // IEEE Trans. Pattern Anal. Mach. Intell. 2009. No. 31(1). P. 39–58.

23. *Maaten L., Hinton G.* Visualizing data using t-SNE // Journal of machine learning research. 2008. Vol. 9. No. Nov. P. 2579–2605.

24. *Chomboon K., Chujai P., Teerarassamee P., Kerdprasop K., Kerdprasop N.* An empirical study of distance metrics for k-nearest neighbor algorithm // Proceedings of the 3rd international conference on industrial application engineering. 2015. P. 280–285.

25. *Vapnik V. N., Chervonenkis A. Ya.* Theory of pattern recognition // Moscow: Nauka, 1974. 416 p.

26. *Vapnik V.* The Nature of Statistical Learning Theory // Springer Science & Business Media 1999, 314 p.

27. *Peterson L. E.* K-nearest neighbor // Scholarpedia. 2009. Vol. 4. No. (2). P. 1883.

28. *Tax D. M., Duin P.* Feature scaling in support vector data descriptions // Learning from Imbalanced Datasets. 2000. P. 25–30.

29. *Busso C., Bulut M., Lee C., Kazemzadeh A., Mower Provost E., Kim S., Chang J., Lee S., Narayanan S.* IEMOCAP: Interactive emotional dyadic motion capture database // Language resources and evaluation. 2008. Vol. 42. No. 4. P. 335.

30. *Cao H., Cooper D., Keutmann M., Gur R., Nenkova A., Verma R.* CREMA-D: Crowd-sourced emotional multimodal actors dataset // IEEE transactions on affective computing. 2014. Vol. 5. No. 4. P. 377–390.

31. *Jackson P., Haq S.* Surrey audio-visual expressed emotion (savee) database // University of Surrey: Guildford, UK, 2014.

32. *Pichora-Fuller M. K., Dupuis K.* Toronto emotional speech set (TESS) // Scholars Portal Dataverse. 2020.

33. *Ekman P.* Basic emotions // In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion. Chichester: Wiley, 1999.

34. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J.* Scikit-learn: Machine learning in Python // The Journal of machine Learning research. 2011. Vol. 12. P. 2825–2830.

35. *Poličar P. G., Stražar M., Zupan B.* Embedding to Reference t-SNE Space Addresses Batch Effects in Single-Cell Classification // In: Kralj Novak P., Šmuc T., Džeroski S. (eds) Discovery Science. DS 2019. Lecture Notes in Computer Science. Springer, Cham, 2019.Vol. 11828.

Uzdiaev Mikhail Yu. — research assistant, Laboratory of Big Data and Socio-Cyberphysical Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences.
E-mail: m.y.uzdiaev@gmail.com
ORCID: https://orcid.org/0000-0002-7032-0291

Ryabinov Artem V. — software engineer, Laboratory of Autonomous Robotic Systems, St. Petersburg Federal Research Centre of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences.
E-mail: iamryabinov@gmail.com
ORCID: https://orcid.org/0000-0002-3572-4493