ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ / INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/IRJ.2024.147.30

ЭФФЕКТИВНОСТЬ ПРИМЕНЕНИЯ РАЗЛИЧНЫХ СПЕКТРАЛЬНЫХ ПРИЗНАКОВ ДЛЯ КЛАССИФИКАЦИИ ЭМОЦИЙ С ПОМОЩЬЮ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

Научная статья

Соболь Б.В.[1], Васильев П.В.[2], Байков С.Э.[3,] *, Гофман Е.В.[4]
[1, 2, 3, 4] Донской государственный технический университет, Ростов-на-Дону, Российская Федерация

* Корреспондирующий автор (st.rnd[at]mail.ru)

Аннотация

В последние годы классификация эмоций в разговорной речи привлекла значительное внимание благодаря её применению в виртуальных ассистентах, обучении и анализе настроений. Несмотря на успехи в англоязычных исследованиях, русскоязычные данные, такие как Dusha и RESD, остаются недостаточно изученными. В этом исследовании анализируются спектральные признаки (MFCC, мел-спектрограмма, хромаграмма, спектральный контраст) для классификации эмоций с использованием сверточной нейронной сети. Эксперименты на наборах данных показали наибольшую точность при использовании мел-спектрограмм. Для набора данных RAVDESS точность составила 78%, для Dusha 62%, для RESD 73%. Комбинация признаков не улучшила результаты. Современные методы, такие как самообучение и трансформеры, эффективны, но требовательны к ресурсам. Предложена упрощенная нейросетевая модель для устройств с ограниченной производительностью, что расширяет её применение на смартфоны, умные часы и системы умного дома, обеспечивая высокую точность при низком энергопотреблении.

Ключевые слова: классификация эмоций, мел-частотные кепстральные коэффициенты, мел-спектрограмма, хромограмма, спектральный контраст, нейросетевой классификатор, самообучение, архитектура трансформеров, русскоязычные данные, искусственный интеллект, машинное обучение.

EFFECTIVENESS OF USING DIFFERENT SPECTRAL FEATURES FOR EMOTION CLASSIFICATION USING A CONVOLUTIONAL NEURAL NETWORK

Research article

Sobol B.V.[1], Vasilev P.V.[2], Baikov S.E.[3,] *, Gofman Y.V.[4]
[1, 2, 3, 4] Don State Technical University, Rostov-on-Don, Russian Federation

* Corresponding author (st.rnd[at]mail.ru)

Abstract

Emotion classification in spoken language has attracted considerable attention in recent years due to its application in virtual assistants, training and sentiment analysis. Despite the successes in English-language studies, Russian-language data such as Dusha and RESD remain understudied. This research analyses spectral features (MFCC, fine spectrogram, chromagram, spectral contrast) for emotion classification using convolutional neural network. Experiments on the datasets showed the highest accuracy using fine spectrograms. For the RAVDESS dataset, the accuracy was 78%, for Dusha 62%, and for RESD 73%. Combining features did not improve the results. Current methods such as self-learning and transformers are efficient but resource demanding. A simplified neural network model for performance constrained devices is proposed, which extends its application to smartphones, smartwatches and smart home systems, providing high accuracy with low power consumption.

Keywords: emotion classification, fine-frequency cepstral coefficients, fine spectrogram, chromogram, spectral contrast, neural network classifier, self-learning, transformer architecture, Russian-language data, artificial intelligence, machine learning.

Введение

В последние годы классификация эмоций в разговорной речи получила значительное внимание со стороны исследователей, поскольку она находит широкое применение в различных областях, включая виртуальных ассистентов, системы обучения и анализа потребительских настроений. Несмотря на существенные успехи в этой области, большинство исследований проводилось на англоязычных данных, что оставляет заметный пробел в применении этих технологий к русскоязычным данным.

Анализ и интерпретация эмоций является сложной задачей [1] как для человека, так и для компьютера. Поскольку речь является основным средством общения между людьми, системы аффективных вычислений, способные распознавать эмоции, занимают важное место в развитии взаимодействия человека и компьютера. Однако точное определение эмоций представляет значительную сложность из-за их многообразной природы и способов выражения.

Эмоциональное состояние человека предоставляет важные данные о его здоровье и психическом благополучии. Например, изменения в эмоциональном состоянии, вызванные болезнью Паркинсона, могут быть выявлены через анализ мимики, речи и ЭЭГ-сигналов [2], [3], [4], [5].

Развитие искусственного интеллекта стимулирует создание систем, моделирующих человеческое поведение, но многие современные виртуальные помощники игнорируют эмоциональные аспекты, полагаясь только на транскрипцию речи.

Большая часть исследований в области распознавания эмоций основана на анализе эмоциональной составляющей по расшифрованному тексту, что неудивительно, так как анализ настроений в письменном виде сопровождается значительно большим объемом работ [6] по сравнению с необработанным аудио. Исследователи достигли высокой точности на таких наборах данных, как Yelp и IMDb. К сожалению, транскрипция не учитывает множество характеристик, присутствующих в аудио.

Для развития систем классификации эмоций в русскоязычном сегменте были созданы специализированные наборы данных, такие как Dusha и RESD, которые позволяют исследователям проводить эксперименты и разрабатывать алгоритмы для распознавания эмоций в русскоязычной речи. Набор данных Dusha, разработанный компанией SberDevices, является на данный момент крупнейшим ресурсом такого рода на русском языке [15]. Он состоит из двух частей: первая часть, названная Crowd, включает тексты, созданные на основе реальных диалогов с виртуальным ассистентом, которые затем озвучивались. Вторая часть, Podcast, содержит короткие фрагменты из русскоязычных подкастов, которые были классифицированы по эмоциональным категориям. Весь датасет включает около 300 тысяч аудиозаписей общей продолжительностью примерно 350 часов и охватывает пять классов эмоций: гнев, грусть, позитив, нейтральная эмоция и прочее.

Другой значимый набор данных, Russian EmotionalSpeechDialogs (RESD), представлен в открытой библиотеке Aniemore и предназначен для анализа эмоциональной окраски разговорной и письменной речи [16]. В этом датасете записи диалогов с заранее заданными эмоциями были озвучены профессиональными актёрами. Он включает около 1400 аудиозаписей общей продолжительностью примерно 4 часа и классифицирует эмоции на семь категорий: нейтральная, гнев, энтузиазм, страх, грусть, радость и отвращение.

Также был использован популярный набор данных Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). В этот набор входят 1440 аудиофайлов, записанных с участием 24 профессиональных актеров (12 мужчин и 12 женщин), которые произносили два одинаковых высказывания с нейтральным североамериканским акцентом. Каждый актер воспроизводил фразы, выражая восемь различных эмоций с разной степенью интенсивности: спокойствие, радость, печаль, гнев, страх, удивление, отвращение и нейтральность. Этот набор данных был сбалансирован, выделяя шесть основных эмоций: гнев, отвращение, страх, радость, печаль и нейтральность [7]. Количество аудиофайлов составило 1152, что соответствует среднему количеству файлов в 196 для каждого класса эмоций. Также аудиофайлы были приведены к частоте дискретизации 16кГц.

Обзор современных технологий классификации эмоций

В последние годы значительное внимание привлекает метод самообучения (Self-Supervised Learning, SSL), ставший ведущим в обработке данных. Этот метод позволяет алгоритмам автоматически обнаруживать закономерности в исходных данных, значительно улучшая результаты анализа. Важную роль в успехе SSL сыграло развитие архитектуры трансформеров, которая, в отличие от традиционных нейронных сетей, обеспечивает параллельную обработку больших объемов данных, ускоряя процесс обучения. Это позволило использовать огромные объемы интернет-данных для создания моделей, которые демонстрируют высокую эффективность даже с минимальной доработкой на специализированных наборах данных [17].

Одной из самых известных моделей самообучения, созданных на основе архитектуры трансформеров, является BERT, разработанный компанией Google для обработки письменного текста. BERT широко используется в различных приложениях, таких как обработка запросов в поисковых системах. Однако для задач, связанных с обработкой звуковой речи, BERT не столь эффективен. Это связано с необходимостью предварительной токенизации звуковых данных, что может вносить шум и снижать качество анализа. Для обработки разговорной речи были разработаны специализированные модели, такие как Wav2vec 2.0, HuBERT и WavLM [18], [19], [20]. Эти модели работают напрямую с необработанными аудиозаписями, разбивая их на короткие фрагменты и обучаясь предсказывать скрытые части записи, что позволяет получать качественные представления аудиоданных без необходимости их предварительной разметки.

Тем не менее, авторы данной работы предлагают упрощенную версию нейросетевого классификатора, рассчитанного на внедрение в устройства с ограниченной производительностью.

Это существенно расширит спектр устройств для применения данной модели, например:

- устройства, такие как смартфоны, умные часы и фитнес-браслеты, часто имеют ограниченные вычислительные ресурсы. Упрощенная нейросетевая модель может быть эффективно использована для распознавания эмоций на таких устройствах, улучшая пользовательский опыт без значительного увеличения энергопотребления;
- в системе умного дома или умного офиса устройства с ограниченной вычислительной мощностью могут использовать данный классификатор для распознавания эмоций и адаптации окружающей среды (например, освещения или температуры) в зависимости от настроения пользователя;
- в образовательных приложениях, таких как интерактивные учебники или виртуальные помощники, распознавание эмоций может улучшить взаимодействие с учащимися, предоставляя более персонализированную помощь. В медицинских приложениях такая модель может помочь в мониторинге психоэмоционального состояния пациентов.

Если проводить сравнение предлагаемого подхода с существующими методами классификации эмоциональной составляющей в аудиоданных, то можно выделить следующее:

- трансформеры (например, BERT, Wav2vec 2.0) обеспечивают высокую точность за счет обработки больших объемов данных параллельно. Однако они требуют значительных вычислительных ресурсов и памяти, что ограничивает их применение на устройствах с ограниченными ресурсами;
- классические модели машинного обучения, такие как SVM и деревья решений, также могут быть применены для распознавания эмоций, но они часто требуют тщательной настройки и предварительной обработки данных.

Упрощенные нейросетевые модели могут предложить более высокую гибкость и способность к обучению на разнообразных данных;

- глубокие нейронные сети (DNN) предлагают высокую точность и способность обрабатывать сложные данные, но они также требуют значительных вычислительных ресурсов.

В целом, упрощенная версия нейросетевого классификатора предлагает компромисс между точностью и вычислительной эффективностью. Упрощенные модели могут быть легко адаптированы и переобучены на новых данных, что позволяет быстро интегрировать их в новые приложения или устройства. За счет снижения требований к вычислительным ресурсам, предложенная модель может быть внедрена в устройства с ограниченной производительностью, не снижая при этом качество распознавания эмоций. Уменьшение сложности модели способствует снижению энергопотребления, что особенно важно для носимых устройств и IoT. Модель может быть адаптирована для различных задач и настроек, что делает ее универсальным инструментом для распознавания эмоций в различных приложениях.

Анализ спектральных признаков и их визуализация

В контексте задачи распознавания эмоций по аудиоданным можно выделить несколько спектральных признаков, используемых для анализа. Одними из наиболее распространенных являются мел-спектрограммы и мел-частотные кепстральные коэффициенты (MFCC). В более редких случаях используются хромограммы и спектральный контраст. Цель данного исследования состоит в сравнении эффективности этих признаков при использовании их для обучения сверточной нейронной сети (CNN).

Мел-спектрограмма представляет собой графическое изображение частотного спектра сигнала, изменяющегося во времени, которое широко применяется для анализа речи [8], [9]. Каждый участок сигнала представлен вертикальной линией на графике, отображающей амплитуду в зависимости от частоты в конкретный момент времени. Цветовая шкала используется для отображения амплитуды частот. График мел-спектрограммы аудио-фрагмента из применяемого набора данных представлен на рисунке 1 слева.

MFCC представляет собой кратковременное спектральное представление мощности звука, основанное на линейном косинусном преобразовании логарифмического спектра мощности, измеренного по нелинейной шкале частот мела. График MFCC представляет собой визуализацию временной зависимости коэффициентов MFCC, используемых для анализа аудиосигнала. На горизонтальной оси отображается время, на вертикальной оси – номера коэффициентов MFCC. Используя цветовую шкалу, график показывает интенсивность каждого коэффициента в децибелах: чем ярче цвет, тем выше значение коэффициента, отражающее магнитуду коэффициента. Спектральные признаки MFCC представлены на рисунке 1 справа.

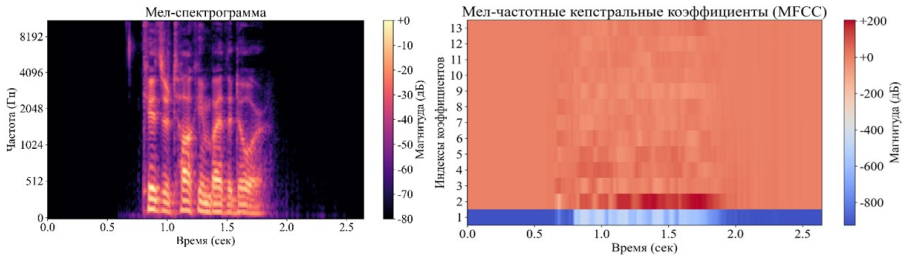

Рисунок 1 - Мел-спектрограмма аудиосигнала (*слева*), MFCC аудиосигнала (*справа*)

DOI: https://doi.org/10.60797/IRJ.2024.147.30.1

Хромаграмма представляет собой визуализацию музыкального сигнала, отображающую интенсивность каждого из 12 хроматических тонов музыкальной октавы во времени. Этот инструмент широко используется в анализе музыкальных данных, поскольку он отражает гармоническую и мелодическую структуру аудиосигнала. В данном исследовании применяются следующие методы построения хромаграмм [10]:

1. Метод, использующий кратковременное преобразование Фурье (STFT) для расчета хромаграммы, который обеспечивает высокое временное разрешение. Это полезно для анализа быстропеременных музыкальных структур. Хромаграмма, построенная с использованием STFT, позволяет детально рассматривать изменения в интенсивности нот во времени. Назовем его ChromaSTFT.

2. Метод, основанный на постоянной Q-преобразовании (Constant-Q Transform, CQT), который позволяет лучше захватывать музыкальные свойства сигнала за счет логарифмически равномерного распределения частотных полос, соответствующих музыкальным нотам. Это делает его особенно полезным для анализа гармонических и мелодических содержаний. Назовем его ChromaCQT.

3. Метод, который использует вариативное Q-преобразование (Variable-Q Transform, VQT), являющееся обобщением постоянной Q-преобразования (CQT). VQT позволяет динамически изменять разрешение в зависимости от частоты, обеспечивая более гибкое и точное представление частотных компонентов музыкального сигнала. Назовем его ChromaVQT.

4. Метод, который использует энергонезависимые хроматические признаки (Chroma Energy NormalizedStatistics, CENS). Этот метод включает дополнительные шаги нормализации и сглаживания, что делает его устойчивым к динамическим изменениям в аудиосигнале и более надежным для задач классификации и сравнения музыкальных фрагментов. Назовем его ChromaCENS.

На графике хромаграммы ось абсцисс представляет время в секундах, ось ординат обозначает 12 хроматических нотили классов высоты тона, цветовая шкала указывает интенсивность каждой ноты. Графики STFT и CENS приведены на рисунке 2. Графики CQT и VQT, приведены на рисунке 3.

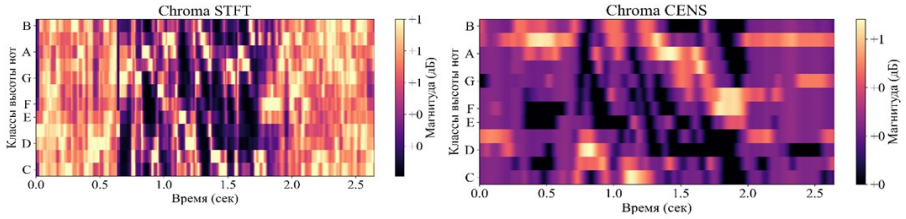

Рисунок 2 - Графики STFT, CENS для аудиосигнала
DOI: https://doi.org/10.60797/IRJ.2024.147.30.2

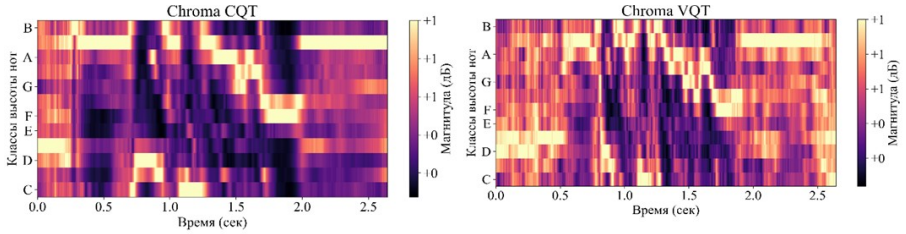

Рисунок 3 - Графики CQT, VQT для аудиосигнала
DOI: https://doi.org/10.60797/IRJ.2024.147.30.3

Спектральный контраст представляет собой разницу между амплитудами пиков и впадин звуковой энергии в частотном спектре [11]. Этот признак особенно полезен для различения музыкальных инструментов, голосов и определения эмоциональных состояний в речи [12]. График спектрального контраста обычно представляет собой матрицу, где по оси абсцисс откладывается время, а по оси ординат – номера частотных полос. Каждая ячейка матрицы отображает значение спектрального контраста для соответствующего временного окна и частотной полосы. Цветовая шкала показывает значение спектрального контраста. Более яркие цвета указывают на более высокий контраст. График спектрального контраста представлен на рисунке 4.

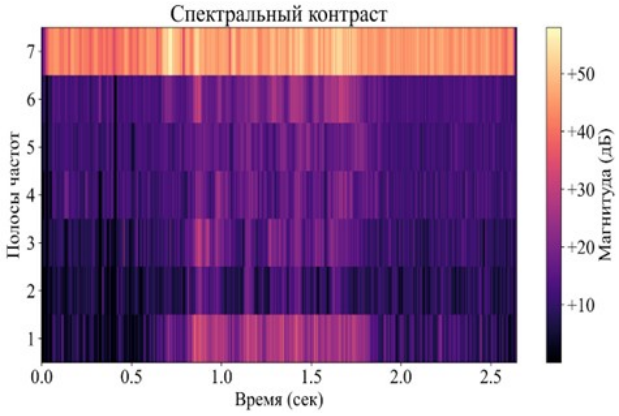

Рисунок 4 - Спектральный контраст аудиосигнала
DOI: https://doi.org/10.60797/IRJ.2024.147.30.4

Модель сверточного классификатора

Для проведения классификации эмоциональной окраски аудио фрагментов на основе извлеченных признаков, была разработана модель сверточной нейронной сети, которая представлена на рисунке 5.

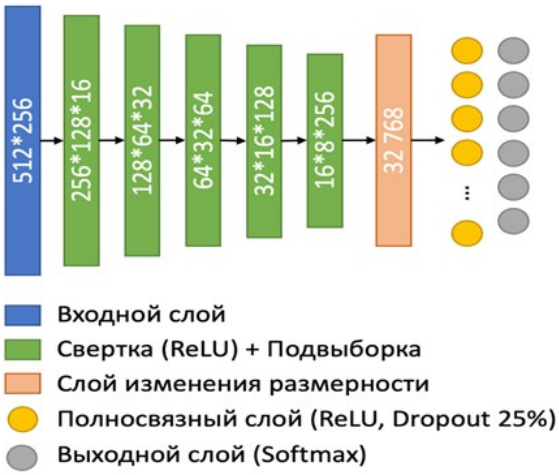

Рисунок 5 - Модель сверточного классификатора
DOI: https://doi.org/10.60797/IRJ.2024.147.30.5

Входной слой данной модели принимает изображения размером 256 на 512 пикселей с одним каналом для черно-белых изображений.

Сверточные слои выполняют функцию извлечения основных характеристик из входного изображения путем применения фильтров. Сверточные операции помогают выделить важные особенности, такие как края, углы и текстуры. Размер фильтров составляет 3x3, с использованием шага (stride) 1 для перемещения фильтра по изображению и padding 1 для сохранения размеров карты признаков.

Слои подвыборки выполняют функцию уменьшения размеров карт признаков для снижения вычислительной сложности и предотвращения переобучения. Они используют операции максимальной подвыборки с окном, 2x2, для уменьшения размерности карт признаков.

Полносвязные слои преобразуют двумерные карты признаков в один вектор признаков для последующей классификации. Каждый нейрон в полносвязном слое соединен с каждым нейроном предыдущего слоя, что позволяет объединить все выявленные признаки и выполнить классификацию.

Слой регуляризации (Dropout Layer) предотвращает переобучение за счет случайного отключения некоторых нейронов во время обучения. Определенный процент нейронов (25%) отключается на каждой итерации обучения, что способствует обобщению модели.

Выходной слой осуществляет классификацию входного изображения по одному из возможных классов. Применяется функция активации softmax для получения вероятностей принадлежности к каждому классу. Количество нейронов на выходном слое равно количеству классов для классификации (в данном случае, 6 классов эмоций).

Результаты исследований

Для оценки эффективности классификации эмоций с использованием различных спектральных признаков применялись изображения размером 256 на 512 пикселей. Нейронная сеть обучалась на каждом спектральном признаке в течение 200 эпох с размером пакета 32. Результаты классификации, достигнутые каждой моделью, приведены в таблице 1.

Таблица 1 - Результаты распознавания эмоций с применением различных спектральных признаков

DOI: https://doi.org/10.60797/IRJ.2024.147.30.6

Спектральный признак	RAVDESS, %	RESD, %	Dusha, %
Мел-спектрограмма	78	73	62
MFCC	75	70	58
Chroma STFT	55	50	44
Chroma CQT	55	50	44
Chroma VQT	60	57	49
Chroma CENS	51	48	39
Спектральный контраст	52	49	37

Анализ данных выявил, что наилучшие результаты были достигнуты при использовании мел-спектрограмм и MFCC, демонстрируя точность 78% и 75% соответственно, что подтверждается в статье [13]. Среди хромаграмм наибольшую точность показала Chroma VQT, достигнув 60%.

В работе [14] авторы отмечают положительное влияние использования мел-спектрограмм с MFCC и других спектральных признаков для улучшения классификации эмоций. В данном исследовании комбинация различных признаков путем добавления их в модель в качестве отдельных каналов не дала значительного прироста точности. Наибольшая точность для комбинаций признаков (69%) была достигнута при использовании изображений мел-спектрограмм, MFCC и Chroma VQT, что не превосходит результаты, полученные при использовании только мел-спектрограмм или MFCC. Похожей точности достигла модель, которая дополнительно содержала изображения спектрального контраста. Результаты классификации с использованием комбинации признаков, достигнутые каждой моделью, приведены в таблице 2.

Таблица 2 - Результаты распознавания эмоций с применением комбинаций спектральных признаков

DOI: https://doi.org/10.60797/IRJ.2024.147.30.7

Спектральный признак	RAVDESS, %	RESD, %	Dusha, %
Мел-спектрограмма, MFCC	65	61	48
Мел-спектрограмма, MFCC, ChromaVQT	69	65	51
Мел-спектрограмма, MFCC, ChromaVQT, Спектральный контраст	68	63	49

Заключение

Мел-спектрограммы и MFCC продемонстрировали наилучшие результаты в классификации эмоций, что подтверждает их значимость и надежность в анализе аудиосигналов. Среди хромаграмм, Chroma VQT показала высшую точность, указывая на её потенциальную ценность в данной области. Использование нескольких каналов для добавления различных спектральных признаков в модель не привело к значительным улучшениям. Точность комбинаций оказалась ниже, чем при использовании отдельных мел-спектрограмм или MFCC, что может указывать на необходимость более тщательного подбора и обработки признаков для повышения эффективности.

Исследование подчеркивает критическую важность выбора правильных спектральных признаков для распознавания эмоций, демонстрируя, что эффективность комбинирования признаков может варьироваться в зависимости от конкретных условий и методов обработки данных, несмотря на их потенциальные преимущества.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Milner R. A Cross-Corpus Study on Speech Emotion Recognition / R. Milner, M.D. Jalal, R.W.M. Ng [et al.]. — 2019. — DOI: 10.1109/ASRU46091.2019.9003838.

2. Dar N.M. EEG-based emotion charting for Parkinson's disease patients using Convolutional Recurrent Neural Networks and cross dataset learning / N.M. Dar, M.U. Akram, R. Yuvaraj [et al.] // Computers in Biology and Medicine. — 2022. — № 144. — DOI: 105327.10.1016/j.compbiomed.2022.105327.

3. Murugappan M. Tunable Q wavelet transform based emotion classification in Parkinson’s disease using Electroencephalography / M. Murugappan, W. Alshuaib, A.K. Bourisly [et al.] // PLoS ONE. — 2020. — № 15 (11). — DOI: 10.1371/journal.pone.0242014.

4. Righi S. Automatic and controlled attentional orienting toward emotional faces in patients with Parkinson's disease / S. Righi, G. Gronchi, S. Ramat [et al.] // Cognitive, affective & behavioral neuroscience. — 2023. — DOI: 23.10.3758/s13415.023.01069.5.

5. Skibinska J. Parkinson’s Disease Detection based on Changes of Emotions during Speech / J. Skibinska, R. Burget. — 2020. — DOI: 10.1109/ICUMT51630.2020.9222446.

6. Yang Z. XLNet: Generalized Autoregressive Pretraining for Language Understanding / Z. Yang, Z. Dai, Y. Yiming [et al.]. — 2019. — DOI: 10.48550/arXiv.1906.08237.

7. 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. — № 13. — DOI: e0196391.10.1371/journal.pone.0196391.

8. Deller J.R. Discrete-Time Processing of Speech Signals / J.R. Deller, H.L.J. Hansen, J.G. Proakis. — MacMillan Pub, 1999. — DOI: 10.1109/9780470544402.

9. Flanagan J.L. Speech Analysis Synthesis and Perception / J.L. Flanagan. — Springer-Verlag Berlin Heidelberg, 1972. — DOI: 10.1007/978.3.662.01562.9.

10. Shah A.K. Chroma Feature Extraction / A.K. Shah, M. Kattel, A. Nepal [et al.] // Chroma Feature Extraction using Fourier Transform. — 2019.

11. Nandini C.S. Modulation spectra of natural sounds and ethological theories of auditory processing / C.S. Nandini, F.E. Theunissen // The Journal of the Acoustical Society of America. — 2003. — № 114. — P. 3394-3411. — DOI: 10.1121/1.1624067.

12. Taffeta M.E. The Modulation Transfer Function for Speech Intelligibility / M.E. Taffeta, F.E. Theunissen // PLoS Computational Biology. — 2009. — № 5. — DOI: 10.1371/journal.pcbi.1000302.

13. Zielonka M. Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets / M. Zielonka, A. Piastowski, A. Czyżewski [et al.] // Electronics. — 2022. — № 11. — P. 3831. — DOI: 10.3390/electronics11223831.

14. Dias I. Speech emotion recognition with deep convolutional neural networks / I. Dias, M.F. Demirci, A. Yazici // Biomedical Signal Processing and Control. — 2020. — Vol. 59. — DOI: 10.1016/j.bspc.2020.101894.

15. Kondratenko V. Large Raw Emotional Dataset with Aggregation Mechanism / V. Kondratenko, A. Sokolov, N. Karpov [et al.]. — 2022. — DOI: 10.48550/arXiv.2212.12266.

16. Давидчук Н. ANIEMORE Открытая библиотека распознавания эмоций в речи человека / Н. Давидчук, И. Любенец, А. Аментес. — 2023. — DOI: 10.13140/RG.2.2.10999.80802.

17. Saeed A. Multi-task Self-Supervised Learning for Human Activity Detection / A. Saeed, T. Ozcelebi, J. Lukkien // Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. — 2019. — Vol. 3. — № 61. — P. 1-30. — DOI: 10.1145/3328932.

18. Schneider S. Unsupervised Pre-Training for Speech Recognition / S. Schneider, A. Baevski, R. Collobert [et al.] // International Speech Communication Association. — 2019. — P. 3465-3469. — DOI: 10.21437/Interspeech.2019.1873.

19. Hsu W.N. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units // W.N. Hsu, B. Bolte, Y.H.H. Tsai [et al.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2021. — P. 1. — DOI: 10.1109/TASLP.2021.3122291.

20. Chen S. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing / S. Chen, C. Wang, Z. Chen [et al.] // IEEE Journal of Selected Topics in Signal Processing. — Vol. 16. — 2022. — P. 1-14. — DOI: 10.1109/JSTSP.2022.3188113.

Список литературы на английском языке / References in English

1. Milner R. A Cross-Corpus Study on Speech Emotion Recognition / R. Milner, M.D. Jalal, R.W.M. Ng [et al.]. — 2019. — DOI: 10.1109/ASRU46091.2019.9003838.

2. Dar N.M. EEG-based emotion charting for Parkinson's disease patients using Convolutional Recurrent Neural Networks and cross dataset learning / N.M. Dar, M.U. Akram, R. Yuvaraj [et al.] // Computers in Biology and Medicine. — 2022. — № 144. — DOI: 105327.10.1016/j.compbiomed.2022.105327.

3. Murugappan M. Tunable Q wavelet transform based emotion classification in Parkinson’s disease using Electroencephalography / M. Murugappan, W. Alshuaib, A.K. Bourisly [et al.] // PLoS ONE. — 2020. — № 15 (11). — DOI: 10.1371/journal.pone.0242014.

4. Righi S. Automatic and controlled attentional orienting toward emotional faces in patients with Parkinson's disease / S. Righi, G. Gronchi, S. Ramat [et al.] // Cognitive, affective & behavioral neuroscience. — 2023. — DOI: 23.10.3758/s13415.023.01069.5.

5. Skibinska J. Parkinson’s Disease Detection based on Changes of Emotions during Speech / J. Skibinska, R. Burget. — 2020. — DOI: 10.1109/ICUMT51630.2020.9222446.

6. Yang Z. XLNet: Generalized Autoregressive Pretraining for Language Understanding / Z. Yang, Z. Dai, Y. Yiming [et al.]. — 2019. — DOI: 10.48550/arXiv.1906.08237.

7. 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. — № 13. — DOI: e0196391.10.1371/journal.pone.0196391.

8. Deller J.R. Discrete-Time Processing of Speech Signals / J.R. Deller, H.L.J. Hansen, J.G. Proakis. — MacMillan Pub, 1999. — DOI: 10.1109/9780470544402.

9. Flanagan J.L. Speech Analysis Synthesis and Perception / J.L. Flanagan. — Springer-Verlag Berlin Heidelberg, 1972. — DOI: 10.1007/978.3.662.01562.9.

10. Shah A.K. Chroma Feature Extraction / A.K. Shah, M. Kattel, A. Nepal [et al.] // Chroma Feature Extraction using Fourier Transform. — 2019.

11. Nandini C.S. Modulation spectra of natural sounds and ethological theories of auditory processing / C.S. Nandini, F.E. Theunissen // The Journal of the Acoustical Society of America. — 2003. — № 114. — P. 3394-3411. — DOI: 10.1121/1.1624067.

12. Taffeta M.E. The Modulation Transfer Function for Speech Intelligibility / M.E. Taffeta, F.E. Theunissen // PLoS Computational Biology. — 2009. — № 5. — DOI: 10.1371/journal.pcbi.1000302.

13. Zielonka M. Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets / M. Zielonka, A. Piastowski, A. Czyżewski [et al.] // Electronics. — 2022. — № 11. — P. 3831. — DOI: 10.3390/electronics11223831.

14. Dias I. Speech emotion recognition with deep convolutional neural networks / I. Dias, M.F. Demirci, A. Yazici // Biomedical Signal Processing and Control. — 2020. — Vol. 59. — DOI: 10.1016/j.bspc.2020.101894.

15. Kondratenko V. Large Raw Emotional Dataset with Aggregation Mechanism / V. Kondratenko, A. Sokolov, N. Karpov [et al.]. — 2022. — DOI: 10.48550/arXiv.2212.12266.

16. Davidchuk N. ANIEMORE Otkrytaja biblioteka raspoznavanija jemocij v rechi cheloveka [Open library for emotion recognition in human speech] / N. Davidchuk, I. Lyubenets, A. Amentes. — 2023. — DOI: 10.13140/RG.2.2.10999.80802. [in Russian]

17. Saeed A. Multi-task Self-Supervised Learning for Human Activity Detection / A. Saeed, T. Ozcelebi, J. Lukkien // Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. — 2019. — Vol. 3. — № 61. — P. 1-30. — DOI: 10.1145/3328932.

18. Schneider S. Unsupervised Pre-Training for Speech Recognition / S. Schneider, A. Baevski, R. Collobert [et al.] // International Speech Communication Association. — 2019. — P. 3465-3469. — DOI: 10.21437/Interspeech.2019.1873.

19. Hsu W.N. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units // W.N. Hsu, B. Bolte, Y.H.H. Tsai [et al.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2021. — P. 1. — DOI: 10.1109/TASLP.2021.3122291.

20. Chen S. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing / S. Chen, C. Wang, Z. Chen [et al.] // IEEE Journal of Selected Topics in Signal Processing. — Vol. 16. — 2022. — P. 1-14. — DOI: 10.1109/JSTSP.2022.3188113.