http://dx.doi.org/10.35596/1729-7648-XXXX-XX-X-XX-XX

*Оригинальная статья*
*Original paper*

УДК *004.934.2+534.784*

# РАСПОЗНАВАНИЕ РЕЧЕВЫХ ЭМОЦИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ОПОРНЫХ ВЕКТОРОВ И НАДСЕГМЕТНЫХ АККУСТИЧЕСКИХ ПРИЗНАКОВ

КРАСНОПРОШИН Д.В. ВАШКЕВИЧ М.И.

*Белорусский государственный университет информатики и радиоэлектроники*
*(г. Минск, Республика Беларусь)*

*Поступила в редакцию*



**Аннотация.** В данном исследовании изучается проблема распознавания речевых эмоций с использованием мел-частотных кепстральных коэффициентов (МЧКК) при помощи классификатора на основе метода опорных векторов (МОВ). В качестве набора данных был использован датасет RAVDESS. Была предложена модель, которая использует 306-компонентный супрасегментарный вектор признаков МЧКК в качестве входных данных для классификатора МОВ. Для оценки качества модели использовался невзвешенное среднее значение полноты (UAR). Эксперименты проводились с различными функции ядра для МОВ (например, линейный, полиномиальный и радиальный базис) и разным размером кадра для извлечения МЧКК (от 20 до 170 мс). Результаты экспериментов демонстрируют многообещающую точность (UAR = 48%), демонстрируя потенциал этого подхода для таких приложений, как голосовые помощники, виртуальные агенты и диагностика психического здоровья.

**Ключевые слова:** голосовой сигнал, МЧКК, ЦОС, извлечение аудио признаков, распознавание, машинное обучение.

**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов.

**Для цитирования.** Вашкевич М.И., Краснопрошин Д.В. Распознавание речевых эмоций с использованием метода опорных векторов и надсегментных акустических признаков. Доклады БГУИР. 2023; **(*): ***-***.

# SPEECH EMOTION RECOGNITION USING SVM CLASSIFIER WITH SUPRASEGMENTAL MFCC FEATURES

MAXIM.I. VASHKEVICH, DANIIL V. KRASNOPROSHIN

*Belarusian state university of informatics and radioelectronics*
*P.Brovki str., 6, Minsk, 220013, Republic of Belarus*

*Submitted*



**Abstract.** This study explores speech emotion recognition (SER) using mel-frequency cepstral coefficients (MFCCs) and Support Vector Machines (SVMs) classifier on the RAVDESS dataset. We proposed a model which uses 306-component suprasegmental MFCC feature vector as an input downstream by SVM classifier. To evaluate the quality of the model, unweighted average recall (UAR) was used. We evaluate different kernel function for SVM (such as linear, polynomial and radial basis) and different frame size for MFCC extraction (from 20 to 170 ms). Experimental results demonstrate promising accuracy (UAR = 48%), showcasing the potential of this approach for applications like voice assistants, virtual agents, and mental health diagnostics.

**Keywords:** voice signal, MFCC, DSP, audio feature extraction, recognition, machine learning.

**Conflict of interests.** The authors declare no conflict of interests.

**For citation.** Vashkevich M.I., Krasnoproshin D.V., Speech emotion recognition using SVM classifier with suprasegment MFCC features. Doklady BGUIR. 2023; **(*): ***-***.

## Введение

Область распознавания эмоций по речи быстро развивается в последние десятилетия благодаря росту производительности вычислительных систем и широкому интересу к этому вопросу исследователей в области психологии, психиатрии и информатики [1], [2]. Эмоции часто влияют на процессы принятия решений, поэтому распознавание эмоций может представлять интерес для построения более эффективного общения, включая диалоговые системы (голосовые помощники, чат-боты). Задача распознавания негативных эмоций, таких как стресс, гнев, усталость является важным аспектом с точки зрения обеспечения безопасности дорожного движения при использовании интеллектуальных транспортных средств, поскольку позволяет им реагировать на эмоциональное состояние водителя [3].

В данной работе рассматривается задаче определения эмоций на основе анализа речевого сигнала. Одна из основных проблем данного подхода связана с определением набора признаков, эффективно описывающих эмоциональное стояние диктора [1], [4-6].

В данной работе для построения системы распознавания эмоций предлагается использовать мел-частотные кепстральные коэффициенты (МЧКК) [7] для получения признаков и метод опорных векторов (МОВ) [7] в качестве метода классификации.

Признаки на основе МЧКК широко применяются в системах распознавания эмоций по речи [1, 5]. МЧКК воспроизводят реакцию слуховой системы человека на звук, улавливая соответствующую акустическую информацию [8]. Формируя представление аудиосигнал в частотной области, МЧКК уменьшают размерность данных, сохраняя при этом важные функции, что делает его пригодным для алгоритмов машинного обучения, таких как МОВ. Также, МЧКК устойчивы к шуму и вариациям стилей речи, гарантируя сохранение тонких эмоциональных нюансов в речи.

В свою очередь МОВ является простым и надежным подходом к задаче классификации, который обеспечивает адаптируемость к многомерным пространствам признаков. МОВ основан на принципе поиска оптимальной гиперплоскости, максимально разделяющей разные классы в пространстве признаков [9]. В контексте распознавания эмоций по речи это означает, что МОВ может эффективно дифференциировать эмоциональные состояния [5]. Кроме того, МОВ может учитывать нелинейные отношения с помощью функций ядра (*kernel function*), что позволяет улавливать сложные закономерности в речевых данных.

## Извлечение речевых признаков

Первым этапом системы по распознаванию эмоций по речи является предварительная обработка входных аудиоданных [1, 5]. В данной работе речевые признаки рассчитывались на

основании МЧКК [8]. Расчет МЧКК относится к методам кратковременного анализа речевого сигнала, которые предполагают разбиение сигнала на фреймы (короткие сегменты). Считается, что в интервале от 10 до 30 мс голосовой сигнал можно считать стационарным. На рис. 1 представлена схема вычисления МЧКК.

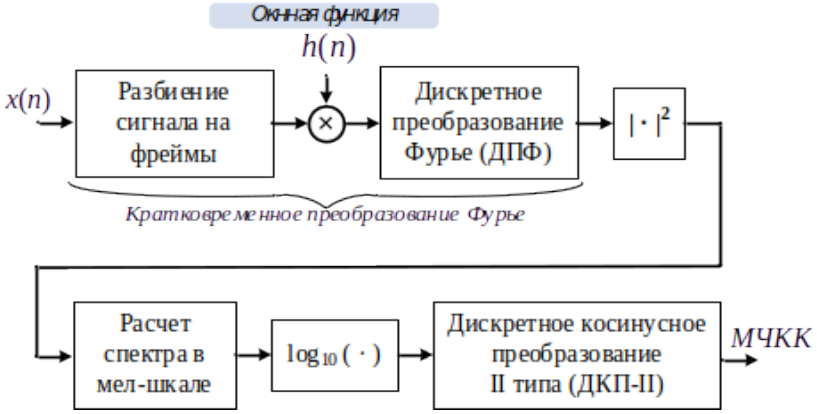



**Рис. 1.** Схема вычисления мел-частотных кепстральных коэффициентов (МЧКК)
**Fig. 1.** Scheme for calculating mel-frequency cepstral coefficients (MFCC)

Согласно рис. 1 процесс извлечения МЧКК включает следующие шаги:

1) вычисление кратковременного преобразования Фурье (КВПФ) и нахождение квадрата модуля КВПФ для получения спектрограммы сигнала;

2) вычисление мел-спектрограммы (энергия сигнала из шкалы герц переводится в мел-шкалу, отражающую свойства человеческого слуха);

3) взятие логарифма от энергии сигнала в мел-частотных полосах;

4) применение декоррелирующего преобразования, в качестве которого используется дискретное косинусное преобразование второго типа (ДКП-II).

В качестве иллюстрации на рис. 2 показан пример речевого сигнала, выражающего эмоцию гнева.

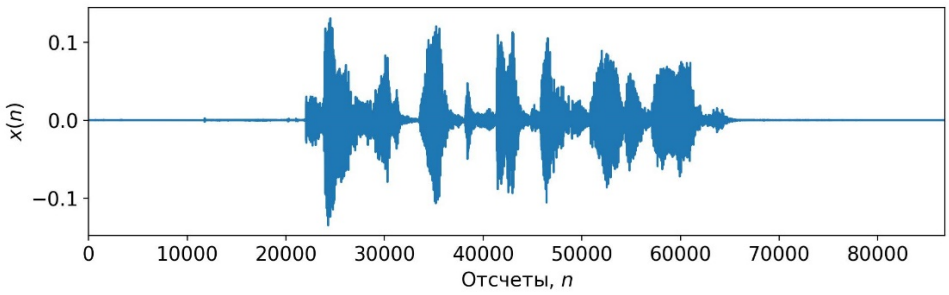
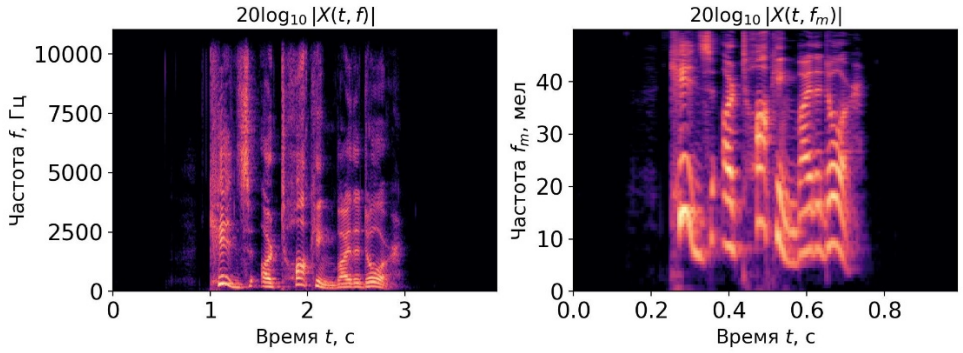



Рис. 2. Представление речевого сигнала, выражающего гнев («Kids are talking by the door»)
Fig. 2. Representation of the speech signal expressing anger

На рис. 3 представлен результат вычисления КВПФ и мел-спектрограммы сигнала, представленного на рис. 1.




а) б)

Рис. 3. речевого сигнала, выражающего гнев: а) спектрограмма (КВПФ), б) мелспектрограмма

Fig. 3. Speech signal expressing anger: a) spectrogram (STFT), b) mel-spectrogramm

На рис. 4 показана временная последовательность MFCC, рассчитанная для сигнала, представленного на рис. 1.

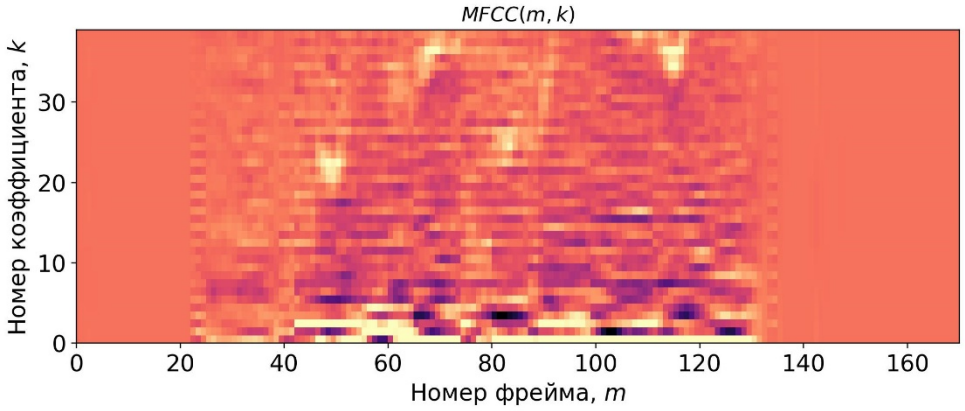



Рис. 4. Временная последовательность МЧКК

Fig. 4. Time-sequence of MFCC

В данной работе используются речевые сигналы с частотой дискретизации 48 кГц. Для обработки аудиосигналов использовалась библиотека librosa написанная на языке Python. КВПФ рассчитывается с использованием следующего набора размеров фреймов $N=\{1024,2048,4096,8192\}$. Размер перекрытия между фреймами брался равным 50% от размера фрейма. Из каждого фрейма извлекалось 40 МЧКК. После обработки одного аудиофайла мы получаем матрицу $M$ МЧКК размером $40\times N_{frames}$, где $N_{frames}$ – количество фреймов. Таким образом, $m$-й столбец матрицы представляет собой вектор МЧКК, вычисленный для временного фрейма с номером $m$.

На основании матрицы $M$ рассчитывался вектор признаков фиксированной длины.

Чтобы получить единый вектор признаков для каждого аудиофайла, мы вычисляем средние (34 признака) и стандартные значения (34 признака) для МЧКК в матрице M вдоль оси времени.

Стандартные значения МЧЧК затем используются для расчета среднеквадратического отклонения (СКО) МЧКК. Это может быть полезно, так как СКО может использоваться для анализа изменчивости или для оценки степени изменения в различных сегментах звукового сигнала.

Более того, к вектору признаков также добавляются первая и вторая производные от стандартного МЧКК. В данном контексте, извлечение первой и второй производных (первой и второй разности) из MFCC коэффициентов имеет физический смысл и помогает в анализе и классификации аудиосигналов.

*Первая производная (первая разность):*

Первая производная MFCC коэффициентов представляет собой скорость изменения каждого MFCC коэффициента во времени. Физический смысл первой производной может быть ассоциирован с изменением спектральных характеристик звука во времени. Например, она может помочь в выявлении моментов, когда звук становится более высокочастотным или более громким, что может быть полезным для распознавания звуковых событий.

*Вторая производная (вторая разность):*

Вторая производная MFCC коэффициентов представляет собой ускорение изменения каждого MFCC коэффициента во времени. Физический смысл второй производной может быть связан с изменением ускорения звука. Например, это может помочь выявить моменты, когда звук начинает быстро увеличиваться или уменьшаться в частоте.

Применение производных MFCC коэффициентов может улучшить способность системы распознавания речи или звукового анализа в обнаружении и классификации различных аудиосигналов. Они могут использоваться для выделения важных характеристик аудиосигнала, таких как изменения в тональности, интонации, и ритме речи, что делает их полезными в приложениях, таких как распознавание речи, детектирование звуковых событий, и музыкальный анализ.

В целом, извлечение производных MFCC коэффициентов позволяет внести в анализ аудиосигнала информацию о его динамике и изменениях во времени, что может улучшить способность системы обработки звука распознавать и классифицировать различные звуковые события.

Помимо МЧКК, их первой и второй производных, а также их стандартных отклонений были рассчитаны следующие статистические показатели (с использвание):

**1) Коэффициент асимметрии (*skewness*).**

Это мера степени асимметрии распределения случайной величины. Она показывает, насколько сильно и в какую сторону смещено распределение относительно своего среднего значения.

Существует несколько разных способов вычислить ассиметрию, но самым часто используемым является формула моментов:

$$skewness = \left(\frac{1}{n}\right) * \sum\left(\frac{\left(X_i - X\right)^3}{s^3}\right)$$

где $X_i$ - значения случайной величины,

$\bar{X}$ - среднее значение,

n - количество наблюдений,

s - стандартное отклонение.

Значение skewness позволяет определить, является ли распределение симметричным (skewness близка к 0) или асимметричным (skewness отличается от 0). Если skewness положительна, то распределение смещено вправо (большинство значений находится в левой части графика), а если skewness отрицательна, то распределение смещено влево (большинство значений находится в правой части графика).

В контексте мел-частотных кепстральных коэффициентов (MFCC) асимметрия может предоставить информацию о форме распределения коэффициентов и подчеркнуть определенные аспекты вариабельности звукового сигнала.

В контексте задач распознавания эмоций в речи это может быть полезным по нескольким причинам:

а) *Отражение эмоциональных особенностей: к*оэффициент асимметрии может отразить, насколько сильно эмоциональные состояния влияют на интонации в речи. Высокая асимметрия может указывать на ярко выраженные и отличительные интонации, что может быть связано с определенными эмоциями.

б) *Анализ изменчивости интонаций:* коэффициент асимметрии может также служить индикатором степени изменчивости в интонациях. Высокая асимметрия может указывать на более выраженные и неоднородные изменения в интонациях, что может быть связано с более динамичными эмоциональными состояниями.

в) *Различение эмоциональных состояний:* различные эмоциональные состояния могут проявляться в различных формах асимметрии в распределении интонаций. Анализ коэффициента асимметрии может помочь выявить характерные особенности распределения для каждой эмоции.

г) *Определение аномалий*: аномальные эмоциональные состояния могут проявляться в отклонениях от нормальной симметрии. Высокая асимметрия может указывать на наличие аномалий в интонациях, которые могут быть связаны с необычными эмоциональными выражениями.

**2) Эксцесс (Kurtosis).**

Это мера формы распределения случайной величины, которая показывает, насколько оно остроконечное или плоское по сравнению с нормальным распределением.

Существуют несколько разных способов вычислить эксцесс, но самым распространенным является формула моментов:

$$kurtosis = \left(\frac{1}{n}\right) * \sum \left(\frac{\left(X_i - X\right)^4}{\left(s^4 - 3\right)}\right)$$

где $X_i$ - значения случайной величины,

$\bar{X}$ - среднее значение,

n - количество наблюдений,

s - стандартное отклонение.

Значение эксцесса позволяет определить, насколько остроконечно или плоское распределение. Если эксцесс положительный, то распределение является остроконечным (есть большое количество значений, сосредоточенных вокруг среднего), а если эксцесс отрицательный, то распределение плоское (есть меньше значений вокруг среднего и больше значений в хвостах распределения).

Расчет эксцесса для МЧКК может быть полезным в контексте задач распознавания эмоций в речи по нескольким причинам:

а) *Острота интонаций:* эксцесс может предоставить информацию о форме распределения интонаций в речи. Острый и высокий пик распределения может отражать более выраженные и четкие интонации, что может быть связано с эмоциональными состояниями.

б) *Характер изменчивости*: эксцесс также может указывать на степень изменчивости в интонациях. Высокий эксцесс может означать, что определенные эмоциональные состояния характеризуются более выраженными и сосредоточенными интонациями.

в) *Обнаружение аномалий*: высокий или низкий эксцесс может служить индикатором аномальных эмоциональных состояний, отличающихся от нормального распределения. Это может помочь выделить случаи, когда речь содержит особенно яркие или необычные эмоциональные выражения.

г) *Различение между эмоциями:* разные эмоциональные состояния могут проявляться в различных формах интонаций. Эксцесс может помочь выявить характерные особенности распределения для каждой эмоции.

**3) Interquartile Range (IQR) (Межквантильный размах).**

Это мера разброса данных, которая используется для измерения разницы между верхним и нижним квартилями. Она показывает дисперсию значений в центральном интервале данных.Для вычисления IQR, нужно выполнить следующие шаги:

а) Упорядочите данные по возрастанию.

б) Найдите значение первого квартиля (Q1), которое разделяет нижнюю 25% наблюдений от верхних 75% наблюдений.

в) Найдите значение третьего квартиля (Q3), которое разделяет нижние 75% наблюдений от верхних 25% наблюдений.

г) Вычислите IQR, как разницу между значениями Q3 и Q1:

**IQR = Q3 — Q1.**

IQR используется для определения наличия выбросов в данных. Обычно выбросами считаются значения, которые находятся за пределами интервала Q1 - 1,5 IQR и Q3 + 1,5 IQR.

В контексте задач распознавания эмоций в речи IQR может быть полезен по по нескольким причинам:

а) *Измерение дисперсии в интонациях:* IQR может предоставить информацию о том, насколько сильно варьируются интонации в речи. Большой IQR может свидетельствовать о более разнообразных и динамичных интонациях, что может быть важным при анализе эмоциональных проявлений.

б) *Определение изменчивости в выражении эмоций:* высокая изменчивость в интонациях может отражать эмоциональную насыщенность и разнообразие выражений. IQR может помочь выделить случаи, когда изменения в интонациях особенно сильны и разнообразны.

в) *Выделение особенностей распределения*: IQR также может служить индикатором особенностей в распределении интонаций. Большой IQR может указывать на наличие ярко выраженных пиков и хвостов, что может быть связано с определенными эмоциональными состояниями.

В итоговый набор признаков были включены среднее значение МЧКК (34 признака), среднеквадратичное отклонение МЧКК (34 признака), среднее от первой и второй производных от МЧКК (2 * 34 признака), их среднеквадратическое отклонение (2 * 34), а также коэффициент асимметрии, эксцесс и межквантильный размах (по 34 признака для каждой метрики соотвественно). Таким образом, для каждого аудиофайла мы получаем **306-компонентный вектор супрасегментных признаков MFCC.**

## Речевая база

При проведении исследования в качестве исходного набора данных использовался Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10]. RAVDESS содержит 7356 записей 24 актеров (12 мужчин, 12 женщин). Все актеры произвели 104 различных вокализации, состоящих из 60 устных высказываний и 44 песенных высказывания. Каждая из 104 вокализаций была экспортирована для создания трех отдельных модальных звуковых условиях: аудио-видео (лицо и голос), только видео (лицо, но без голоса) и только аудио (голос, но без лица). На каждого актера приходилось 312 файлов (104 × 3). Записи одного участника были потеряны по техническим причинам (132 файла). Таким образом, 24 × 312–132 = 7356 файлов. Этот набор состоит из 4320 записей речи и 3036 песен. Актеры озвучили две разных фразы (в речи и песни). Две фразы произносились с восемью эмоциональными окрасками (нейтральность, спокойствие, счастье, грусть, злость, страх, удивление и отвращение). В случае с песнями использовалось шесть эмоциональных окрасок (нейтральность, спокойствие, счастье, грусть, злость и страх). Все эмоциональные состояния, кроме нейтрального, озвучивались на двух уровнях эмоциональной громкости (нормальная и повышенная). Актеры повторяли каждую вокализацию дважды.

В рамках данной работы будет использована только часть датасета RAVDESS, а именно RAVDESS Emotional speech audio. Эта часть RAVDESS содержит 1440 файлов в формате wav (16 бит, 48 кГц): 60 записей на каждого из 24-х профессиональных актера (12 мужчин, 12 женщин). Фразы с нейтральным североамериканским акцентом. Речевые эмоции включают выражения нейтральности, спокойствия, счастья, грусти, гнева, страха, удивления и отвращения. Все эмоциональные состояния, кроме нейтрального, озвучивались на двух уровнях эмоциональной громкости (нормальная и повышенная). Актеры повторяли каждую вокализацию дважды.

## Подход к описанию эксперимента

При проведении экспериментов и проверки эффективности МЧКК для решения задачи распознавания эмоций в речи применялся **метод опорных векторов (МОВ)**.

Метод опорных векторов выполняет классификацию путем построения N-мерных гиперплоскостей, которые оптимально разделяют данные на отдельные категории. Классификация достигается путем построения в пространстве входных данных линейной (или нелинейной) разделяющей поверхности. Идея данного подхода заключается в преобразовании (с помощью функции ядра) исходного набора данных в многомерное пространство признаков. И уже в новом пространстве признаков добиться оптимальной в определенном смысле классификации.

В качестве ядра используется любая симметричная, положительно полуопределенная матрица K, которая составлена из скалярных произведений пар векторов $x_i$ и $x_j$, где $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, характеризующих меру их близости, а $\phi$ является произвольной преобразующей функцией, формирующее ядро. В частности, примерами таких функций являются:

- **линейное ядро:**

$$K(x_i, x_j) = x_i^T x_j,$$

что соответствует классификатору на опорных векторах в исходном пространстве

- **полиномиальное ядро со степенью p:**

$$K(x_i, x_j) = (1 + x_i^T x_j,)^p$$

- **гауссово ядро с радиальной базовой функцией (RBF):**

$$K(x_i, x_j) = \exp\left(\gamma \left\| x_i - x_j \right\|^2\right)$$

В качестве ядерной функции модели на основе МОВ была выбрана линейная. Значение параметра C (cost) (допустимый штраф за нарушение границы зазора) было равно единице.

Построение классификатора на опорных векторах с использованием перечисленных выше ядер можно, в частности, осуществить с помощью библиотеки sklearn, написанной на языке Python.

Для тренировки, тестирования и валидации модели использовался метод k-блочной кросс-валидация (k-fold cross-validation) [12].

Метод k-блочной кросс-валидации включает следующие действия:

1) Перемешать набор данных случайным (псевдо-случайным) образом;

2) Разделить набор на k групп;

3) Для каждой уникальной группы:

- выделить группу записей в качестве тестовых данных (test data)
- взять оставшиеся группы в качестве тренировочных данных (train data)
- обучить модель на тренировочных и оценить ее эффективность на тестовых данных
- сохранить значение оценки и сбросить модель до исходного состояния для следующей итерации
- установить средний уровень навыка модели.

В данной работе данных были разбиты на блоки следующим образом (в скобках указаны номера актеров):

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

Такой порядок разбиения был предложен в работе мультимодальному распознаванию эмоций на наборе данных RAVDESS с использованием трансферного обучения [13]. Выбранная стратегия заключается в том, что каждый блок должен содержать одинаковое количество случайно выбранных образцов для каждого класса. При этом должно выполняться условие, что каждый актер представлен либо обучающей, либо валидационной выборке, но не в обоих [13].

Для оценки качества модели было вычислено среднее арифметическое (невзвешенное) полноты (UAR). UAR — это показатель, используемый для измерения общей производительности модели многоклассовой классификации. Он вычисляет средний уровень запоминания по всем классам, придавая каждому классу одинаковую важность без учета классового дисбаланса. Формула невзвешенного среднего отзыва (UAR) определяется следующим образом:

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

где A – матрица путаницы,

$N_c$ – количество классов. Значение $UAR$ находится в диапазоне от 0 до 1.

Эксперимент проводился в три этапа:

1) подготовка обучающей выборки;

2) обучение и тестирование классификатора с использованием другой функции ядра и других параметров анализа речи;

3) оценка модели с использованием метрики $UAR$.

## Результаты и их обсуждение

Эксперименты, проведенные с набором данных RAVDESS с использованием классификаторов SVM с различными ядрами и гиперпараметрами, включая RBF, линейные и полиномиальные ядра, а также с различной длиной кадров для извлечения МЧКК, дали ценную информацию о распознавании эмоций. Мы использовали технику grid search, чтобы настроить и найти лучшие гиперпараметры для данного ядра. В таблице 1 дана краткая информация обо всех проведенных экспериментах.

Таблица 1

Результирующий UAR для классификатора на основа МОВ с различными ядрами

| Размер фрейма | Линейное ядро | Полиномиальное ядро | RBF ядро |
|---|---|---|---|
| 1024 | 0.458 (C =0.01) | 0.457(C = 0.01, γ= 1, deg= 1) | 0.469 (C = 8.11, γ= 0.0008) |
| 2048 | 0.451 (C =0.1) | 0.45 (C = 0.01, γ= 1, deg= 1) | 0.471 (C = 8.11, γ= 0.00088) |
| 4096 | 0.454 (C =0.01) | 0.455 (C = 0.05, γ= 0.1, deg= 1) | 0.476 (C = 2.31, γ= 0.0014 |
| 8192 | **0.469** (C =0.01) | **0.474** (C = 0.05, γ= 0.1, deg= 1) | **0.482** (C = 28.48, γ= 0.014) |

Наилучшее значение UAR 48% достигается при использовании SVM с ядром RBF и супрасегментными функциями MFCC, рассчитанными на основе кадров размером 4096. Поверхность UAR, рассчитанная в ходе поиска по сетке для этой модели, представлена на рис. 5. Видно, что более высокое значение параметров C приводит к более гибкому классификатору с более высокой производительностью.

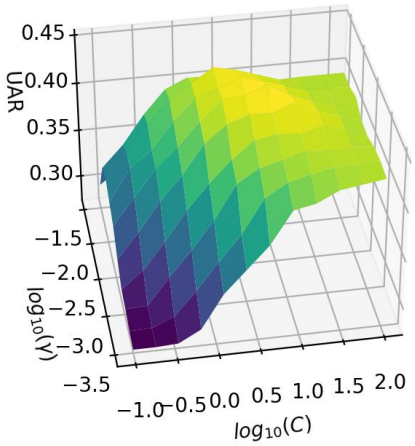



Рис.5. Поверхность UAR

Fig. 5. UAR surface

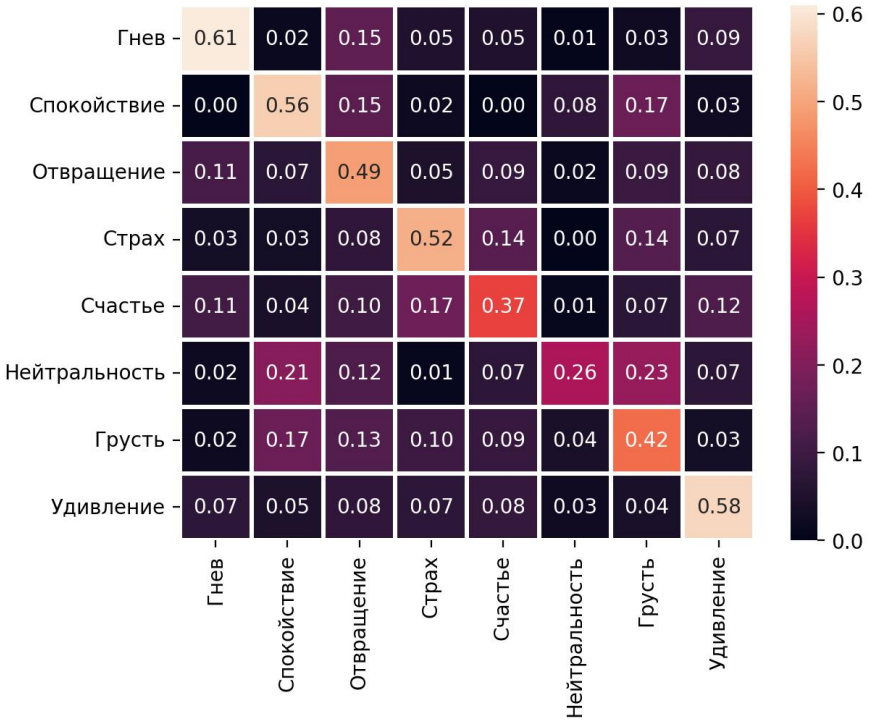



Рис 6. Мультиклассовая матрица спутывания (Multiclass confusion matrix)

Fig 6. Multiclass confusion matrix

На рис. 6 представлена матрица спутывания для лучшей модели SVM-RBF. Анализ матрицы путаницы набора данных RAVDESS с использованием классификатора SVM выявляет важные закономерности в распознавании эмоций. Среди эмоций было замечено, что наиболее часто неправильно классифицированной эмоцией была нейтральность (27%). Интересно, что эту эмоцию часто путают с грустью, что позволяет предположить некоторое сходство их акустических характеристик. И наоборот, «Удивление» продемонстрировало высокую точность распознавания (61%) и редко ошибочно классифицировалось как другая эмоция, что указывает на отличительные особенности его акустического профиля. Эти результаты проливают свет на проблемы, с которыми сталкивается классификатор при

различении тонких эмоциональных нюансов, и подчеркивают важность разработки функций и совершенствования моделей для улучшения эффективности распознавания эмоций.

Наши результаты показывают, что выбор ядра оказывает существенное влияние на точность классификации. Ядро RBF продемонстрировало высокую производительность в отношении множества эмоций, в то время как линейное ядро превосходно различало определенные эмоциональные состояния. Примечательно, что размер кадра, используемый для извлечения MFCC, играл значительную роль в общей точности системы: более короткие кадры обеспечивают более мелкие временные детали, а более длинные кадры собирают более широкую контекстную информацию. Эти результаты подчеркивают важность точной настройки ядра классификатора SVM и учета компромиссов, связанных с размером кадра, при разработке систем распознавания эмоций.

## Заключение

В сфере взаимодействия человека и компьютера точное распознавание эмоций по речи является ключевым фактором. В этой работе представлен подход к проблеме распознавания речевых эмоций, основанный на классификаторе SVM и сверхсегментарных функциях MFCC. Наилучшие результаты (UAR = 48%) получены при использовании SVM-RBF с характеристиками MFCC, рассчитанными на основе кадров длительностью 85 мс. По сравнению с другими работами [2]–[5] есть возможности для улучшения.

## Список литературы / References

1. D. Issa, M. F. Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” Biomedical Signal Processing and Control, vol. 59, 2020.
2. C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, “Multimodal emotion recognition on RAVDESS dataset using transfer learning,” Sensors, vol. 21, no. 22, pp. 1–29, 2021.
3. Xiao, H.; Li, W.; Zeng, G.; Wu, Y.; Xue, J.; Zhang, J.; Li, C.; Guo, G. On-Road Driver Emotion Recognition Using Facial Expression. *Appl. Sci.* 2022, *12*, 807. https://doi.org/10.3390/app12020807
4. S. Sadok, S. Leglaive, and R. Séguier, “A vector quantized masked autoencoder for speech emotion recognition,” arXiv preprint arXiv:2304.11117, 2023.
5. A. Bhavan, P. Chauhan, R. R. Shah et al., “Bagged support vector machines for emotion recognition from speech,” Knowledge-Based Systems, vol. 184, pp. 1–7, 2019.
6. M. Baruah and B. Banerjee, “Speech emotion recognition via generation using an attention-based variational recurrent neural network,” Proc. Interspeech 2022, pp. 4710–4714, 2022.
7. Yu, C., Tian, Q., Cheng, F., Zhang, S. (2011). Speech Emotion Recognition Using Support Vector Machines. In: Shen, G., Huang, X. (eds) Advanced Research on Computer Science and Information Engineering. CSIE 2011. Communications in Computer and Information Science, vol 152. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21402-8_35
8. X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, Spoken language processing: A guide to theory, algorithm, and system development. Prentice hall PTR, 2001.
9. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.
10. C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, “Mel-frequency cepstral coefficient analysis in speech recognition,” in 2006 International Conference on Computing & Informatics, 2006, pp. 1–5.
11. M. M. Goodwin, “The STFT, sinusoidal models, and speech modification,” Springer Handbook of Speech Processing, pp. 229–258, 2008.
12. S. R. Livingstone and F. A. Russo, “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, vol. 13, no. 5, p. e0196391, 2018.
13. Luna-Jiménez, C.; Griol, D.; Callejas, Z.; Kleinlein, R.; Montero, J.M.; Fernández-Martínez, F. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. Sensors 2021, 21, 7665. https://doi.org/10.3390/s21227665

## Вклад авторов

Вашкевич М.И. цель и задачи исследования, предложил идею барк-частотного кепстрального представления голосового сигнала, выполнил программную реализацию расчета БЧКК, принимал участие в подготовке текста статьи и интерпретации результатов экспериментов. Лихачев Д.С. выполнил программную реализацию расчета МЧКК, участвовал в подготовке программной базы для эксперимента.

## Authors contribution

Vashkevich M.I. determined the purpose and objectives of the study, proposed the idea of the bark-frequency cepstral representation of the voice signal, carried out the software implementation of the BFCC calculation, took part in the preparation of the text of the article and the interpretation of the experimental results. Likhachov D.S. carried out the software implementation of the calculation of the MFCC, participated in the preparation of the software tools for the experiment.

**Сведения об авторах**

Краснопрошин Д.В., магистрант кафедры электронных вычислительных средств ФКСиС БГУИР

Вашкевич М.И., д.т.н., профессор кафедры электронных вычислительных средств (ЭВС) Белорусского государственного университета информатики и радиоэлектроники (БГУИР).

**Information about the authors**

D.V. Krasnoproshin master student, Department of Electronic Computing Facilities in BSUIR

M.I. Vashkevich Professor, Department of Electronic Computing Facilities in BSUIR, DrSc.

**Адрес для корреспонденции**

220013, Республика Беларусь, г. Минск, ул. П. Бровки, д. 6, Белорусский государственный университет информатики и радиоэлектроники
тел. +375-17-293-84-78;
e-mail: vashkevich@bsuir.by
Вашкевич Максим Иосифович

**Address for correspondence**

220013, Republic of Belarus, Minsk, P. Brovki str., 6, Belarusian State University of Informatics and Radioelectronics
tel. +375-17-293-84-78;
e-mail: vashkevich@bsuir.by
Vashkevich Maksim Iosifovich