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 СУПРАСЕГМЕНТАРНЫХ ПРИЗНАКОВ МЧКК

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 (SER) 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 80-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. Построение признакового пространства для распознавния эмоций по речевым сигналам. 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 80-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., Feature space construction for speech emotion recognition. Doklady BGUIR. 2021; **(*): ***-***.

Введение

Область компьютерного распознавания речевых эмоций (SER) начала быстро развиваться в последние десятилетия благодаря росту производительности вычислительных ресурсов и широкому интересу исследователей в области неврологии, 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).

Проблема распознавания эмоций в настоящее время является актуальной и прикладной задачей искусственного интеллекта. Ее решение позволяет, например, в сфере связи построить эффективную связь между компьютером и человеком, в сфере медицины (интерфейсы на основе речевых технологий для пользователей с ограниченными возможностями, слепыми или слабовидящими), при принятии решений. задачи (распознавание 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) и т. д.

В данной работе рассматривается подход к решению задачи, основанный на обработке речевых сигналов. При этом одна из основных проблем данного подхода связана с определением набора признаков, эффективно описывающих данный тип эмоций [1], [3]–[5]. Таким образом, происходит построение признакового пространства, в котором могут быть разделены объекты, соответствующие разным классам эмоций.

Для решения столь нетривиальной задачи были использованы два основных метода: извлечение мел-частотных кепстральных коэффициентов (MFCC) в качестве основы для конвейера проектирования признаков и машины опорных векторов (SVM) в качестве алгоритма классификации.

MFCC — широко распространенный и эффективный метод выделения признаков для распознавания речевых эмоций [1], [4]. MFCC воспроизводят реакцию слуховой системы человека на звук, улавливая соответствующую акустическую информацию [6]. Преобразуя аудиосигнал в представление частотной области, MFCC выделяет основные характеристики речи, такие как форма спектра и высота звука. Этот метод уменьшает размерность данных, сохраняя при этом важные функции, что делает его пригодным для алгоритмов машинного обучения, таких как SVM. Более того, MFCC устойчивы к шуму и вариациям стилей речи, гарантируя сохранение тонких эмоциональных нюансов в речи. В результате MFCC служит ценным инструментом в распознавании речевых эмоций, позволяя моделям точно и надежно различать эмоциональные состояния по аудиосигналам.

В то же время SVM предлагает многообещающий подход к распознаванию речевых эмоций, сочетающий в себе надежные возможности классификации с адаптируемостью к многомерным пространствам признаков. SVM основаны на принципе поиска оптимальной гиперплоскости, максимально разделяющей разные классы в пространстве признаков [7]. В контексте распознавания речевых эмоций это означает, что SVM может эффективно различать

различные эмоциональные состояния [4]. Кроме того, SVM может обрабатывать нелинейные отношения с помощью функций ядра, что позволяет им улавливать сложные закономерности в речевых данных. Их способность хорошо обобщать и смягчать переобучение делает SVM подходящим для зачастую шумной и нюансированной эмоциональной речи.

Извлечение речевых признаков

Первым этапом системы по распознаванию речевых является предварительная обработка входных речевых данных [1], [4]. Анализ существующих подходов к категоризации признаков показал, что наиболее подходящей для целей исследования является методика, основанная на расчете МЧКК [6]. Эти признаки широко используются при распознавании эмоций в речи и являются чрезвычайно эффективными инструментами для построения различных моделей машинного обучения [5], [8].

Кепстральное представление голоса в психоакустических шкалах

В данном разделе рассматривается кепстральное представление голосового сигнала, получаемое на основе спектрального анализа сигнала в психоакустически мотивированной частотной шкале. Анализируется широко применяемое для описания голосового сигнала мел-частотное кепстральное представление [6], которое сравнивается с предлагаемым в работе барк-частотным кепстральным представлением, получаемым на основе неравнополосного ДПФ-модулированного банка фильтров.

Расчет мел-частотных кепстральных коэффициентов (МЧКК) относится к методам кратковременного анализа голосового сигнала, которые предполагают разбиение сигнала на кадры анализа. Как правило, в интервале от 10 до 30 мс голосовой сигнал можно считать стационарным. Для больше наглядности предлагается схема вычисления МЧКК показана на рис. 1.

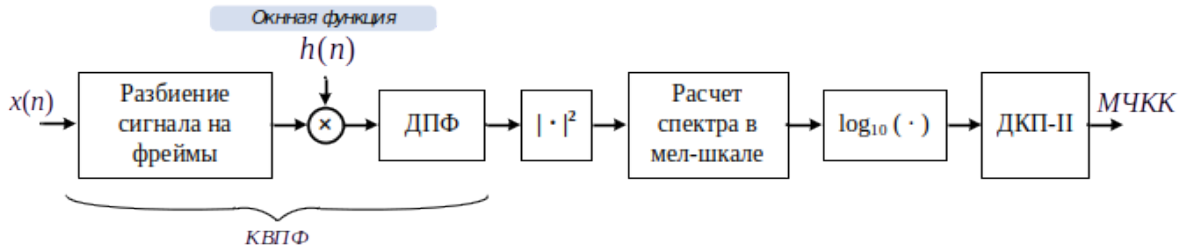

Рис. 1. Схема вычисления мел-частотных кепстральных коэффициентов (МЧКК)
Fig. 1. Scheme for calculating mel-frequency cepstral coefficients (MFCC)

Процесс извлечения Мел-частотных кепстральных коэффициентов включает следующие шаги:

1) **Кратковременное преобразование Фурье (КВПФ):** это особый вид преобразования Фурье, благодаря которому можно узнать, как частоты в сигнале меняются во времени. Он работает, разрезая ваш сигнал на множество небольших сегментов и выполняя преобразование Фурье для каждого из них. В результате обычно получается каскадный график, показывающий зависимость частоты от времени;

Кратковременное преобразование Фурье (КВПФ) (англ. STFT – short-time Fourier *transform*) широко используется для анализа, модификации и синтеза звуковых сигналов [4-5]. КВПФ можно рассматривать как преобразование со скользящим окном, которое имеет вид [5]:

$$X(k,l)=\sum_{n=0}^{N-1} h(n)x(n+lL)e^{-j\omega_k n}$$

где $x(t)$ – входной сигнал, $h(n)$ – ограниченная во времени оконная функция, а $\omega_k = 2\pi k/M$, $k = 0,1,\ldots M-1$ – частотный индекс, L – временной шаг анализа (расстояние между соседними фреймами), l – номер фрейма анализа. Легко заметить, что (1) является вычислением дискретного преобразования Фурье (ДПФ) для сигнала $h(n)x(n+lL)$. Таким образом, представление $X(k,l)$, получающееся в результате КВПФ, является последовательностью локализованных во времени спектров.

На рис. 2 показан пример речевого сигнала из базы данных RAVDESS, а на рис. 3 — спектрограмма (выход КВПФ).

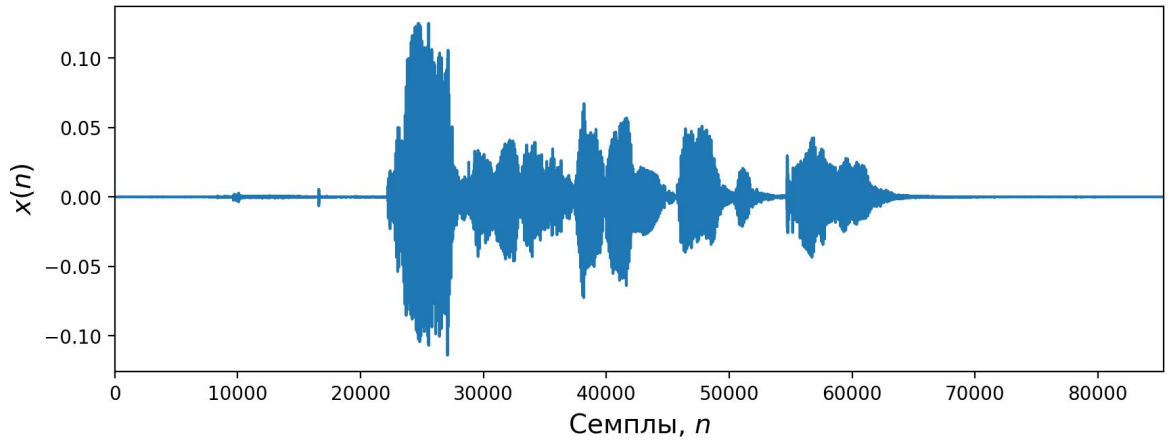

Рис. 2. Представление речевого сигнала, выражающего гнев
Fig. 2. Representation of the speech signal expressing anger

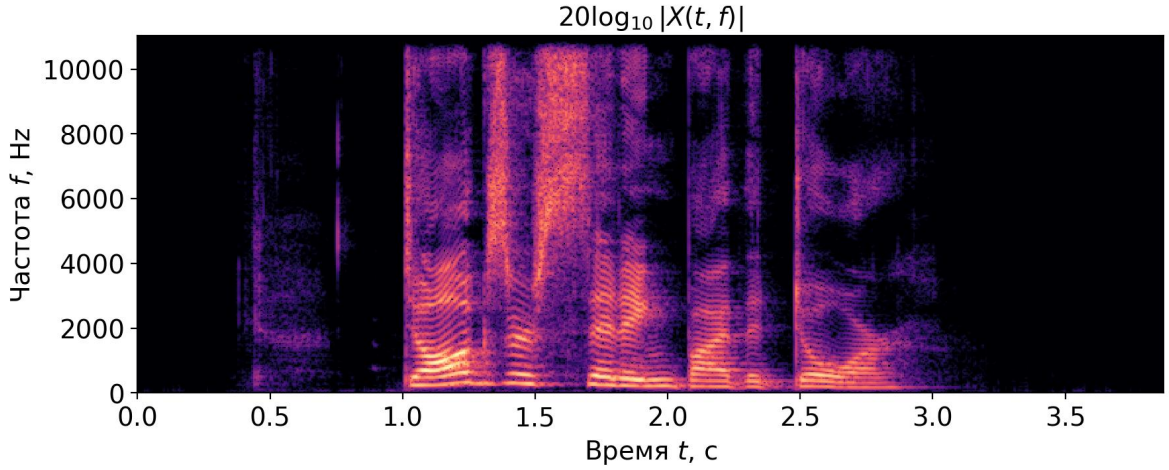

Рис. 3. Спектрограмма речевого сигнала, выражающего гнев
Fig. 3. Spectrogram of a speech signal expressing anger

2) **Расчет набора из М-фильтров:** используется для моделирования свойств человеческого слуха на этапе извлечения признаков. Поэтому мы будем использовать шкалу Мел, чтобы сравнить фактическую частоту с частотой, которую воспринимают люди.

Банк мел-фильтров представляет собой набор треугольных фильтров, равномерно расположенных по шкале мел-частот. Эти фильтры используются для преобразования спектра мощности в мел-частотную область.

Банк мел-фильтров применяется к выходному сигналу КВПФ $X(k,l)^2$ ∨ для получения мел-спектрограммы.

Отметим, что человеческий слух менее чувствителен к изменению энергии звукового сигнала при более высокой энергии по сравнению с более низкой энергией. Логарифмическая функция также имеет аналогичное свойство, при низком значении входного x градиент логарифмической функции будет выше, но при высоком значении входного градиента значение меньше. Поэтому мы применяем log к выходу Mel-фильтра, чтобы имитировать человеческий слух.

3) **Дискретное косинусное преобразование (ДКП)**: Проблема с полученной спектрограммой заключается в том, что коэффициенты банка фильтров сильно коррелированы. Поэтому нам нужно декоррелировать эти коэффициенты. Для этого применяется ДКП.

В результате мы получаем набор чисел, которые являются мел-частотными кепстральными коэффициентами. На рис. 4 показана временная последовательность MFCC, рассчитанная для сигнала, представленного на рис. 2.

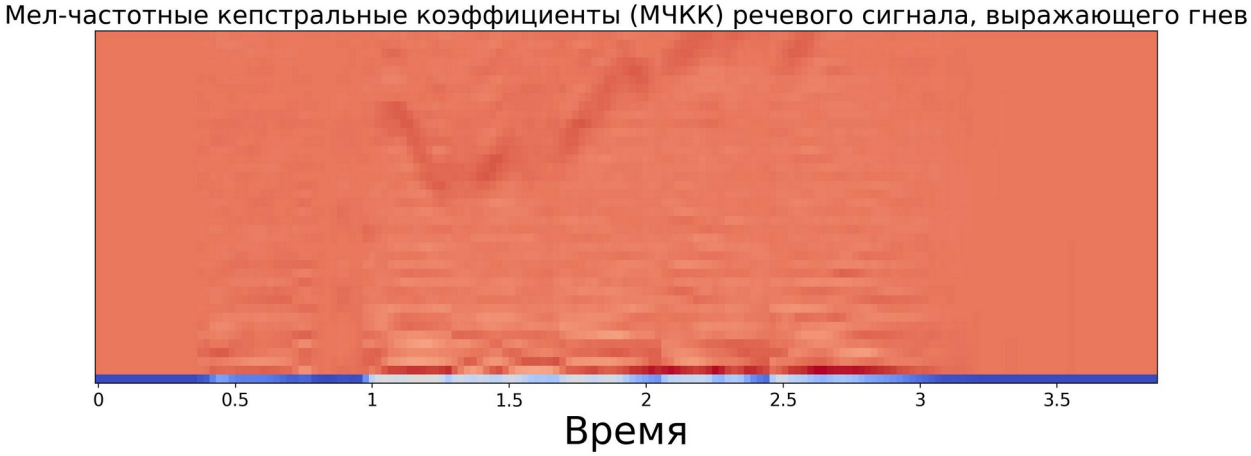

Рис. 4. Временная последовательность МЧКК
Fig. 4. Time-sequence of MFCC

В данной работе используются речевые сигналы с частотой дискретизации 48 кГц. КВПФ рассчитывается с использованием следующего набора размеров кадров N = {1024, 2048, 4096, 8192}. Размер перехода L установлен на N/2. Из каждого кадра N-выборки мы извлекаем 34 МЧКК, используя библиотеку Librosa написанную на языке Python. После обработки одного аудиофайла мы получаем матрицу M МЧКК размером 34 × N_{frames}, где N_{frames} — количество таймфреймов. Чтобы получить единый вектор признаков для каждого аудиофайла, мы вычисляем средние и стандартные значения для МЧКК в матрице M вдоль оси времени.

Более того, к вектору признаков также добавляются первая и вторая производные от 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 были рассчитаны и добавлены к векторов признаков следующие статистические показатели:

1) Skewness (англ. «ассиметрия»).

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:

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

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).

2) Kurtosis (англ. "эксцесс").

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:

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

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).

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:

$$\mathbf{IQR = Q3 — Q1.}$$

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.

Таким образом, для каждого аудиофайла мы получаем 304-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) [9]. 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 ϕ 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)** [10].

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)

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.

Характеристики машины, на которой проводился эксперимент:

1. Процессор AMD Ryzen 7 5700U with Radeon Graphics;
2.Видеокарта AMD Radeon RX Vega 8 (Ryzen 4000/5000) (- 1900 MHz);
3. ОЗУ 16 ГБ DDR4-2400;
4. ОС Ubuntu 20.04.5 LTS;

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.

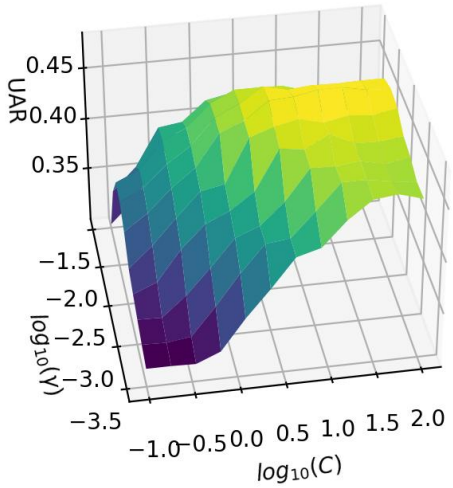

Рис.5. Поверхность UAR
Fig. 5. UAR surface

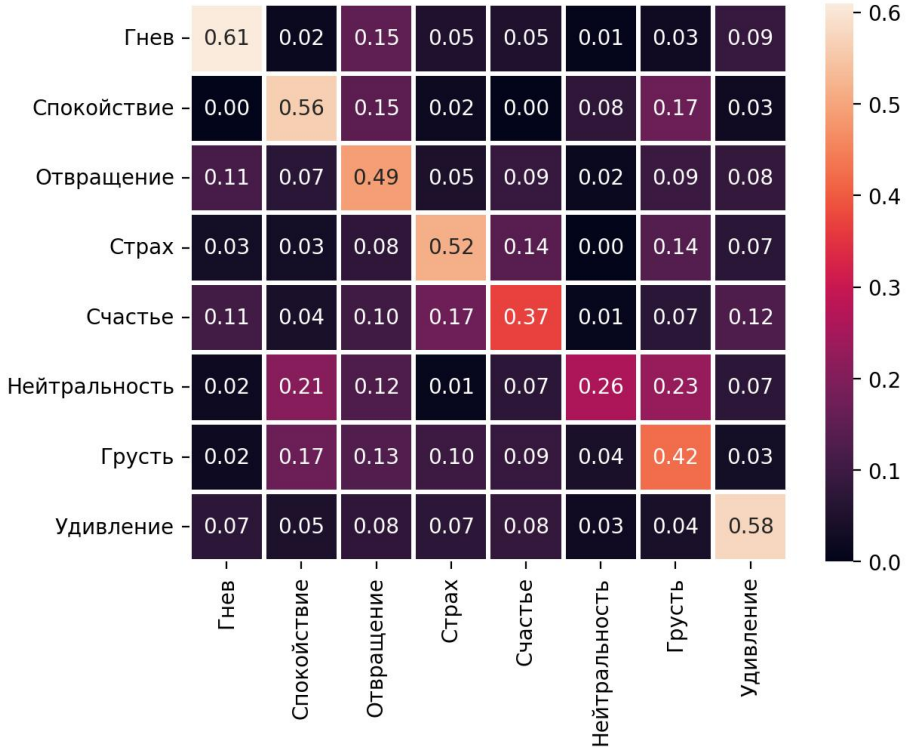

Рис 6. Мультиклассовая матрица спутывания (Multiclass confusion matrix)
Fig 6. Multiclass confusion matrix

На рис. 6 представлена многоклассовая матрица путаницы 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]–[4] 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. S. Sadok, S. Leglaive, and R. Séguier, “A vector quantized masked autoencoder for speech emotion recognition,” arXiv preprint arXiv:2304.11117, 2023.
4. A. Bhavan, P. Chauhan, R. R. Shah et al., “Bagged support vector ma- chines for emotion recognition from speech,” Knowledge-Based Systems, vol. 184, pp. 1–7, 2019.
5. 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.
6. X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, Spoken language processing: A guide to theory, algorithm, and system develop- ment. Prentice hall PTR, 2001.
7. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.
8. 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.
9. M. M. Goodwin, “The STFT, sinusoidal models, and speech modification,” Springer Handbook of Speech Processing, pp. 229–258, 2008.
10. 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.

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. Азаров И.С. предложил идею совместного использования кепстральных признаков и пертурбационных параметров, принимал участие в подготовке текста статьи и интерпретации результатов экспериментов.

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. Azarov I.S. proposed the idea of the joint use of cepstral features with perturbation parameters, took part in the preparation of the text of the article and interpretation of the experimental results.

Сведения об авторах

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

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

Information about the authors

M.I. Vashkevich Professor, Department of Electronic Computing Facilities in BSUIR, PhD of Technical sciences

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

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: sanko@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