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

Оригинальная статья
Original paper

УДК *004.934.2+534.784*

ПОСТРОЕНИЕ ПРИЗНАКОВОГО ПРОСТРАНСТВА ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ В РЕЧИ

КРАСНОПРОШИН Д.В. ВАШКЕВИЧ М.И.

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

Поступила в редакцию

Аннотация. В работе описан алгоритм выделения признаков для задачи классификации эмоций в речевых сигналах

Ключевые слова: голосовой сигнал, МЧКК, БЧКК, 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. 2021; **(*): ***-***.

FEATURE SPACE CONSTRUCTION FOR SPEECH EMOTION RECOGNITION

MAXIM.I. VASHKEVICH, DANIIL V. KRASNOPROSHIN

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

Submitted

Abstract. The paper describes an approach to design a system for analyzing and classification of a voice signal based on perturbation parameters and cepstral presentation. Two variants of the cepstral presentation of the voice signal are considered: based on mel-frequency cepstral coefficients (MFCC) and based on bark-frequency cepstral coefficients (BFCC). The work used a generally accepted approach to calculating the MFCC based on time-frequency analysis by the method of discrete Fourier transform (DFT) with summation of energy in subbands. This method approximates the frequency resolution of human hearing, but has a fixed temporal resolution. As an alternative, a variant of the cepstral presentation based on the BFCC has been proposed. When calculating the BFCC, an warped DFT-modulated filter bank was used, which approximates the frequency and temporal resolution of hearing. The aim of the work was to compare the effectiveness of the use of features

based on the MFCC and BFCC for the designing systems for the analysis and classification of the voice signal. The results of the experiment showed that in the case of using acoustic features based on the MFCC, it is possible to obtain a voice classification system with an average recall of 80.6%, and in the case of using features based on the BFCC, this metric is 83.7%. With the addition of the set of MFCC-features with perturbation parameters of the voice, the average recall of the classification increased to 94.1%, with a similar addition to the set of BFCC-features, the average recall of the classification increased to 96.7%.

Keywords: voice signal, MFCC, BFCC, 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

1) Обозначить актуальность

Изучение эмоций стало быстро развиваться в последнии несколько десятилетий благодаря снижению стоимости вычислительных ресурсов и широкому интерес со стороны исследователей в области неврологии, психологии, психиатрии и информатики. Более того, эмоции зачастую влияют на процессы принятия решений. В связи с этим распознавание эмоций может представлять инетерес, так как знание чувств другого человека позволяет выстраивать более эффективную коммуникацию. Анализируя поведение людей, можно также обнаружить потерю доверия или изменение внутреннего состояния. Это может позволить различным системам, таким как голосовые помощники и чат-боты реагировать на подобные события и адаптировать свои действия для улучшения взаимодействия или изменения содержания диалога, тона или выражения лица, чтобы обеспечить положительный пользовательский опыт.

Речь является одним из основных средств общения между людьми. Она позволяет передать человеческие эмоции и состояние души. В настоящее время предпринимаются попытки реализовать схожий функционал в приложениях, связанных с речью, таких как персональные цифровые помощники, приложения для преобразования текста в речевые модели, сенсоры и др. Исходя из этого, возникает естественная необходимость научить компьютер взаимодействовать так же, как люди, в том смысле, что он мог бы научиться понимать эмоции, лежащие в основе разговорной речи и адекватно реагировать на них.

2) Области применения

В настоящее время обработка речевых сигналов, включая распознавание эмоций, находит применение во множестве сфер. Далее перечислены только некоторые из них:

- здравоохранение;
- распознавание негативных эмоций таких как стресс, злость, усталость является важным аспектом сточки зрения обеспечения безопасности дорожного движения с применением интеллектуальных транспортных средств, поскольку позволяет им реагировать на эмоциональное состояние водителя.
- интерфейсы, созданные на базе речевых технологий для пользователей-инвалидов, слепых или слабовидящих;
- системы компьютерной телефонии, в частности, диалоговых информационно-справочных системах;
- системы управления различными процессами, например, информационные и навигационные системы, диспетчерские системы управления наземным и воздушным транспортом;
- система обработки и защиты речевых сообщений. Одной из функций такой системы является компрессия речи с целью повышения эффективности криптографической защиты

переданного речевого сообщения, а также повышение помехоустойчивости в процессе передачи сообщения по каналу передачи данных;

- системы распознавания речи и идентификации личности, применяющиеся в криминалистической экспертизе, базирующиеся на возможности идентифицировать личность говорящего по голосу;

- системы оценки качества обслуживания в call-центрах и службах поддержки и т. д.

3) Существующие ограничения

Следует отметить, что задача распознавания эмоций, в том числе и в речи является сложной и многомерной, потому что различные эмоции могут быть переданы разным способом и в разных формах.

Для построения схожих с описанными выше систем требуется эффективно решать определенный набор задач. В качестве инструментов для их решения, могут, например, выступать различные модели машинного обучения.

Для построения и обучения таких моделей используются признаки - числовое представление некоторого аспекта сырых данных. Качественные и количественные характеристики признаков играют ключевую роль на протяжении всего процесса создания системы.

В рамках задачи распознавания эмоций в речи возникает дополнительная задача, а именно построение **признакового пространства**. Стоит отметить, что в любой задачи машинного обучения применяются математические модели к данным, чтобы проводить классификацию, делать аналитические выводы или предсказания. Эти модели принимают на вход признаки. **Признак** — это числовое представление некоторого аспекта исходных данных. Признак находится между данными и моделью в процессе машинного обучения. Конструирование признаков — это процесс извлечение некоторых значимых из необработанных данных и приведение их к формату, пригодному для обработки моделью машинного обучения.

Это один из важнейших шагов во всем процессе, так как правильно подобранные признаки облегчают сложное моделирование и, как следствие, способствуют выводу более качественных результатов.

Несмотря на всю важность, отдельно данная тема исследуется недостаточно. Возможно, это происходит потому, что правильные признаки можно определить только в контексте модели и данных, а так как данные и модели могут быть очень разнообразными, сложно выделить общую тактику конструирования признаков для различных проектов.

Важно упомянуть, что дополнительную сложность предоставляет обработка неструктурированных данных. Неструктурированные данные — это наборы данных, которые не были структурированы заранее определенным образом. Неструктурированные данные, как правило, текстовые, такие как открытые ответы на опросы и разговоры в социальных сетях, но также могут быть нетекстовыми, например изображения, видео и аудио.

При этом, подавляющее большинство новых данных, генерируемых сегодня, неструктурировано, что приводит к появлению новых платформ и инструментов, способных управлять ими и анализировать их.

В связи с вышесказанным, большое значение приобретают вопросы, связанные с процессом построения моделей, умеющих эффективно работать с неструктурованными данными. При этом правильно подобранные признаки неотъемлемая часть этого процесса.

В настоящее время рядом специалистов [1–7] уже был выполнен анализ методов и алгоритмов извлечения речевых признаков для построения различных моделей машинного обучения, решающих различные прикладные задачи. Среди недостатков этих работ следует выделить почти полное отсутствие информации о признаках, подходящих для распознавания эмоций в речи. Более того, большинство существующих подходов являются эвристическими.

4) Цели и задачи исследования

Исходя из вышесказанного, авторы данного исследования ставят цель провести анализ существующих подходов и предложить собственный алгоритм выделения признаковых пространств для распознавания эмоций в речи.

Основная часть

Как было упомянуто ранее, для построения системы по распознаванию эмоций в речи требуется провести предобработку исходных данных. Основной задачей предобработки является удаление шума, повышение высоких частот сигнала и получение плоского частотного спектра сигналов, а также частотных характеристик.

Среди проблем связанных с обработкой речи особое место занимает выделение и выбор признаков. Различные аудио признаки позволяют описывать различные аспекты звукового сигнала для решения разного рода прикладных задач.

Существует несколько подходов для категоризации аудио признаков, которые могут варьироваться:

1) С точки зрения уровня абстракции:
- высокоуровневые (инструментовка, аккорды, мелодия, ритм, темп, жанр и т. д.)
- среднеуровневые (дескрипторы, связанные с высотой тона и битами, модели колебаний, мел-кепстральные коэффициенты и др.)
- низкоуровневые (огибающая амплитуда, энергия, спектральный центроид, спектральный поток (spectral flux), спектральный контраст, спектральный спад, спектральная ширина, скорость пересечения нуля (zero crossing rate) и др.)

2) С точки зрения временного охвата:
- мгновенные (~50 мс)
- на уровне отдельных фрагментов (измеряется в секундах)
- глобальные (например, рассматривает отдельно взятую песню целиком)

3) С точки зрения музыкальных аспектов:
- биты
- тембр
- пич (от англ. pitch)

Высота звука без учёта октавы, а точнее множество всех звуковых высот, отстоящих друг от друга на целое число октав.

Качество звука определяется частотой производимых им вибраций; степень высокого или низкого тона.
- благозвучность

4) С точки зрения цифровой обработки сигналов:
- временная область (time domain): огибающая амплитуда, среднеквадратическая энергия, скорость пересечения нуля
- частотная область (frequency domain): отношение полосы частот, спектральный центроид, спектральный поток
- временно-частотная область (time-frequency domain): спектрограммы, мел-спектрограммы, преобразование постоянной Q (constant-Q transform)

5) С точки зрения машинного обучения:
- традиционный подход: огибающая амплитуда, энергия, спектральный центроид, спектральный поток, спектральный контраст, спектральный спад, спектральная ширина, скорость пересечения нуля, отношение полосы частот и т.д.
- подход, базирующийся на использование глубокого обучения: на вход модели подаются не отдельные признаки, а сам сигнал целиком. Модель же сама ищет закономерности и извлекает значимые для нее признаки.

Анализа имеющихся подходов для категоризации признаков показал, что техника на основе расчета Мел-частотных кепстральных коэффициентов является наиболее подходящей для целей исследования. Эти показатели широко используются при распознавания эмоций в речи и являются крайне эффективным инструментов для построения различных моделей машинного обучения.

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.

Процесс извлечения Мел-частотных кепстральных коэффициентов включает следующие шаги:

1) **АЦ-преобразование:** на этом этапе мы преобразуем наш аудиосигнал из аналогового в цифровой формат с частотой дискретизации 22 кГц;

2) **Предыскажение:** увеличивает величину энергии на более высокой частоте. В случаях, когда рассматривается частотная область звукового сигнала для звонких сегментов, таких как гласные, видно, что энергия на более высокой частоте намного меньше, чем энергия на более низких частотах. Повышение энергии на более высоких частотах повысить точность и производительность модели;

3) **Кратковременное преобразование Фурье (STFT):** 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 каждого из них. В результате обычно получается каскадный график, показывающий зависимость частоты от времени;

4) **Расчет набора из М-фильтров:** используется для моделирования свойств человеческого слуха на этапе выделения признаков, что позволяет улучшить производительность модели. Поэтому мы будем использовать шкалу Мела, чтобы сопоставить фактическую частоту с частотой, которую воспринимают люди. Формула отображения приведена ниже:

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.

5) **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 получим набор чисел, являющихся Мел-частотными кепстральными коэффициентами (МЧКК).

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.

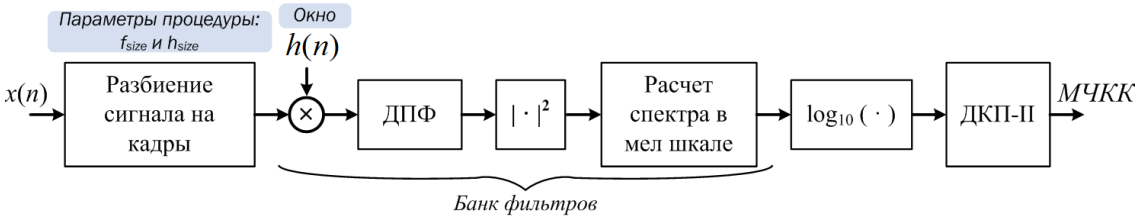

Рис. 1. Схема вычисления мел-частотных кепстральных коэффициентов (МЧКК)
Fig. 1. Scheme for calculating mel-frequency cepstral coefficients (MFCC)

Речевая база

1) Описание RAVDESS

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.

2) Подход к описанию эксперимента (evaluation design)

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 работы модели было вычислено среднее арифметическое (невзвешенное) полноты рассчитанной для каждого распознанного класса.

Полнота представляет собой отношение $ИП/(ИП+ЛО)$, где ИП — количество истинных положительных результатов, а ЛО — количество ложноотрицательных результатов. Также под полнотой понимается интуитивно способность классификатора находить все положительные образцы.

Значение полноты лежит в диапозоне от 0 до 1.

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

Эксперимент проводился в три этапа:

1) подготовка обучающей выборки;

Для большей наглядности будет продемонстрирован процесс вычесления признаков на примере речевого сигнала, выражающего гнев:

а) исходный сигнал (рис. 2)

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

б) Спектрограмма речевого сигнала выражающего гнев (рис. 3)

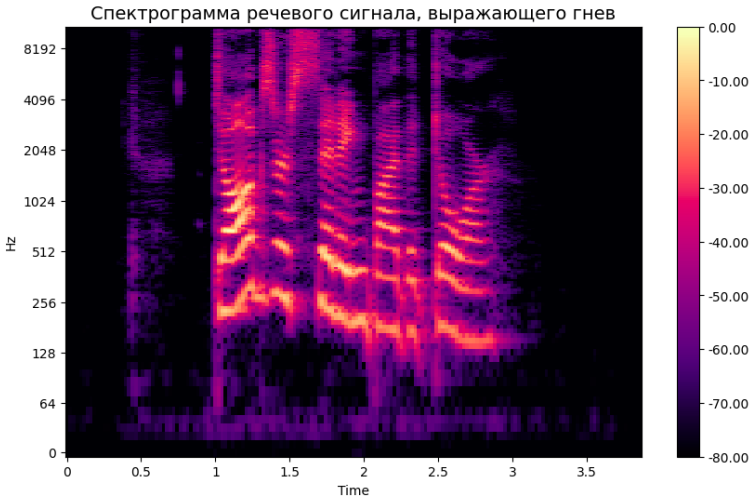

Рис. 3. Спектрограмма речевого сигнала выражающего гнев
Fig. 3. Spectrogram of a speech signal expressing anger

в) Вычисленные по фреймам мел-частотные кепстральные коэффициенты (МЧКК) (рис. 4)

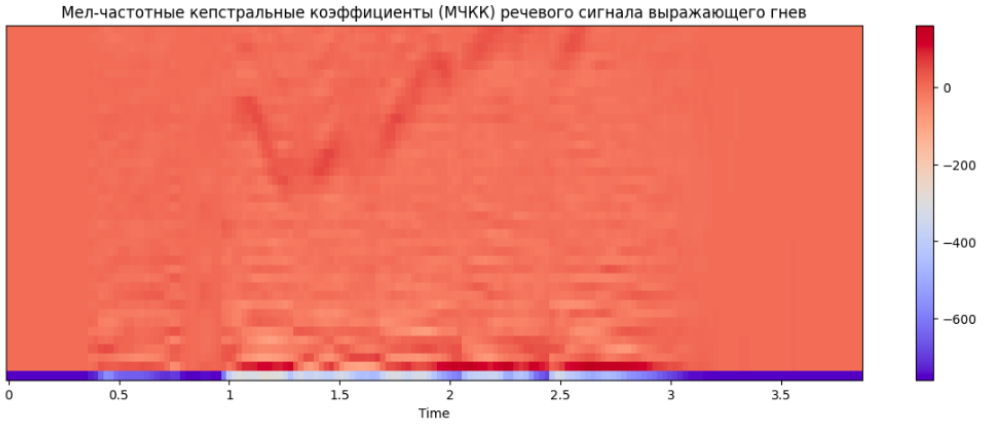

Рис. 4. Вычисленные по фреймам мел-частотные кепстральные коэффициенты (МЧКК)
Fig. 4. Frame-based mel-frequency cepstral coefficients (MCCC)

г) усредненный вектор мел-частотных кепстральных коэффициентов (МЧКК) (рис. 5)

Рис. 5. усредненный вектор мел-частотных кепстральных коэффициентов (МЧКК)
Fig. 5. averaged vector of mel-frequency cepstral coefficients (MFCC)

2) ранжирование признаков;
3) обучение и тестирование классификатора с использованием различного числа признаков.

В результате построения и обучения модели был получен классификатор, точность предсказаний которого при использовании тестового набора данных и вышеуказанной метрики качества достигала 33.7%.

Далее будет представлена мультиклассовая матрица спутывания (англ. Multiclass Confusion Matrix) представляющая собой таблицу или диаграмму, показывающая точность прогнозирования классификатора в отношении двух и более классов. Ячейки таблицы заполняются количеством прогнозов классификатора. Правильные прогнозы идут по главной диагонали от верхнего левого угла в нижний правый.

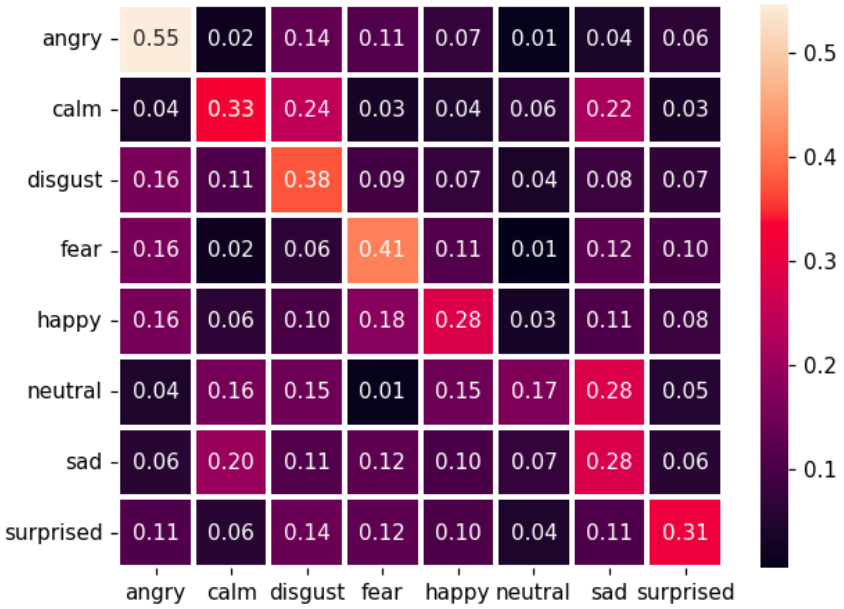

Таблица 1 — Мультиклассовая матрица спутывания (Multiclass confusion matrix)

Полученные в эксперименте результаты суммируются в табл. 3, где приведена оценка средней полноты классификации в зависимости от набора кепстральных признаков, при ограничении на число признаков. Такое ограничение обусловлено тем, что для классификации всегда желательно иметь меньшее число признаков, но обеспечивающих лучшее качество. Лучший результат достигается при использовании набора признаков BFCC-40 (см. табл. 3).

Таблица 2. Максимальная средняя полнота классификации при ограничении числа признаков p.

Table 2. The maximum average recall of the classification achieved when the number of features p is limited.

Набор признаков	Средняя полнота, % (В скобках указано число признаков, при котором достигается средняя полнота)					
	$p \leq 10$	$p \leq 20$	$p \leq 30$	$p \leq 40$	$p \leq 50$	$p \leq 60$
MFCC-10	76,9 (5)	77,2 (15)	–	–	–	–
MFCC-20	61,9 (9)	62,4 (17)	67,6 (30)	70,1 (40)	74,0 (48)	–
MFCC-40	78,0 (9)	80,6 (12)	–	–	–	–
BFCC-10	72,2 (10)	73,5 (11)	–	–	–	–
BFCC-20	77,1 (10)	77,4 (13)	80,0 (28)	–	–	–
BFCC-40	**83,7** (6)	–	–	–	–	–

Заключение

В работе предложен метод вычисления барк-частотных кепстральных коэффициентов (БЧКК), основанный на использовании неравноплосного ДПФ-модулированного банка фильтров, аппроксимирующего частотно-временное разрешение слуха человека. Произведено сравнение предложенных БЧКК с широко распространенными мел-частотными кепстральными коэффициентами (МЧКК) в отношении эффективности построения на их основе системы анализа и классификации голосовых сигналов. Проведенные эксперименты по построению системы классификации голосов пациентов с неврологическим заболеванием БАС показали

эффективность применения надсегментных БЧКК признаков. Среди классификаторов, использующих набор кепстральных признаков лучший результат (средняя полнота 83,7%) достигнут LDA-классификатором, использующим 6 надсегментных БЧКК признаков, отобранных методом LASSO. Среди классификаторов, использующих набор кепстральных признаков, объединенных с пертурбационными параметрами голоса, лучший результат (средняя полнота 96,7%) достигнут LDA-классификатором, использующим 45 надсегментных БЧКК признаков, отобранных методом LASSO.

Список литературы / References

1. Delac, K., Grgic, M., & Grgic, S. Independent comparative study of PCA, ICA, and LDA on the FERET data set. International Journal of Imaging Systems and Technology, 2005, 15(5), 252–260.
2. Pandiyan, “Mel-frequency cepstral coefficient analysis in speech recognition,” Computing & Informatics 2006, ICOCI’06, no. 2, pp. 2–6.
3. L. Xie, Z.-H. Fu, W. Feng, and Y. Luo, “Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news,” Multimedia Systems, vol. 17, pp. 101–112, 2011.
4. Suliman, A., Omarov, B., Dosbayev, Zh. Detection of impulsive sounds in stream of audio signals. 2020 8th International Conference on Information Technology and Multimedia, ICIMU 2020, 2020, pp. 283–287.
5. Rajkumar Palaniappan, K. Sundaraj, «Respiratory Sound Classification using Cepstral Features and Support Vector Machine», 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS). 978-1-4799-2178-2/13.
6. Назаров М. В., Прохоров Ю. Н. Методы цифровой обработки и передачи речевых сигналов. М.: Радио и связь, 1985. 176 с.
7. Сорокин В.Н. Структура проблемы автоматического распознавания речи // Информационные технологии и вычислительные системы, 2004, № 2. С. 25–40.
8. IoT, туман и облака: поговорим про технологии? [Электронный ресурс]. URL: https://3-info.ru/post/2814 (Дата обращения: 21.03.2022).
9. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [Электронный ресурс]. URL: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio?datasetId=107620 (Дата обращения: 21.05.2023).
10. Issa, D.; Fatih Demirci, M.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control 2020, 59, 101894.
11. Luna-Jiménez, Cristina, et al. "Multimodal emotion recognition on ravdess dataset using transfer learning." Sensors 21.22 (2021): 7665.

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

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

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

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

Пертурбационные параметры голоса

Пертурбационные параметры также рассчитываются исходя контура частоты основного тона (ЧОТ). К этой группе относят: 1) частотный диапазон фонации (англ. *PFR – phonatory frequency range*); 2) среднеквадратичное отклонение ЧОТ – SD_{f_o}; 3) энтропия периодов ОТ (англ. *PPE – pitch period entropy*) [4]; 4) индекс патологичности вибрато (англ. *PVI – pathology vibrato index*). Используемые в работе пертурбационные параметры приведены в таблице 1. Более подробное их описание можно найти в работе [5].

Таблица 1. Пертурбационные параметры голоса
Table 1. Perturbation voice parameters

Группа параметров	Число параметров	Названия параметров
Частотная пертурбация	5	$J_{loc}, J_{ppq3}, J_{ppq5}, J_{ppq55}, DPF$
Амплитудная пертурбация	5	$S_{loc}, S_{apq3}, S_{ppq5}, S_{ppq11}, S_{ppq55}$
Пертурбация контура ЧОТ	4	SD_{f0}, PFR, PPE, PVI
Всего	14	