УДК 004.021:004.75

ИЗВЛЕЧЕНИЕ ЗНАЧИМЫХ ХАРАКТЕРИСТИК АУДИОСИГНАЛА ДЛЯ КЛАССИФИКАЦИИ ЭМОЦИЙ В РЕЧИ

М.И. Вашкевич
Профессор кафедры электронных вычислительных средств ФКСиС БГУИР, доктор технических наук
vashkevich@bsuir.by

Д.В. Краснопрошин
Магистрант кафедры электронных вычислительных средств ФКСиС БГУИР
daniil.krasnoproshin@gmail.com

М.И. Вашкевич
Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с исследованием методов и алгоритмов акустического анализа голоса для выявления патологий, синтезом быстрых алгоритмов цифровой обработки сигналов, аппаратной и программной реализацие алгоритмов цифровой обработки сигналов.

Д.В. Краснопрошин
Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с исследованием проблем радиочастотной идентификации объектов, организацией учебного и научно-исследовательского процессов в техническом университете.

Аннотация. TBD

Ключевые слова: ЦОС, извлечение аудио признаков, распознавание, машинное обучение, обработка данных.

Введение.

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.

Эмоции также будут играть важную роль в дальнейших этапах научно-технической революции, которая потребует разработки большего количества социальных роботов. Эти роботы должны будут воспринимать эмоции людей, передавать и создавать свои эмоциональные состояния, чтобы продемонстрировать более тесное личное взаимодействие между человеком и машиной.

Речь является одним из основных средств общения между людьми. Речь позволяет передать свои 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.

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.

В рамках задачи распознавания эмоций в речи возникает дополнительная задача, а именно извлечение признаков. 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. Но, 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 отметить, что отдельную 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. Эти инструменты позволяют организациям с большей легкостью использовать преимущества неструктурированных данных для бизнес-аналитики (BI) и иных прикладных задач.

Большое значение в этой ситуации приобретают вопросы, связанные с процессом построения моделей, умеющих эффективно работать с неструктурованными данными. При этом правильно подобранные признаки неотъемлемая часть этого процесса.

Актуальность.

Технологии обработки речевых сигналов находят применение во множестве сфер. Далее перечислены только некоторые из них:

- 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.

В основе вышеуказанных систем лежат 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.

Рядом специалистов [1–7] 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. В связи с этим в данной работе предлагается изучить типы и процесс извлечения значимых характеристик аудиосигнала для классификации эмоций в речи.

Основная часть.

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.

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.

Как было упомянуто ранее, 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.

В рамках данной работы для извлечения признаков использовалась техника на основе расчета Мел-частотных кепстральных коэффициентов. Эти показатели широко используются при распознавания эмоций в речи и являются крайне эффективным инструментов для построения различных моделей машинного обучения.

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).

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.

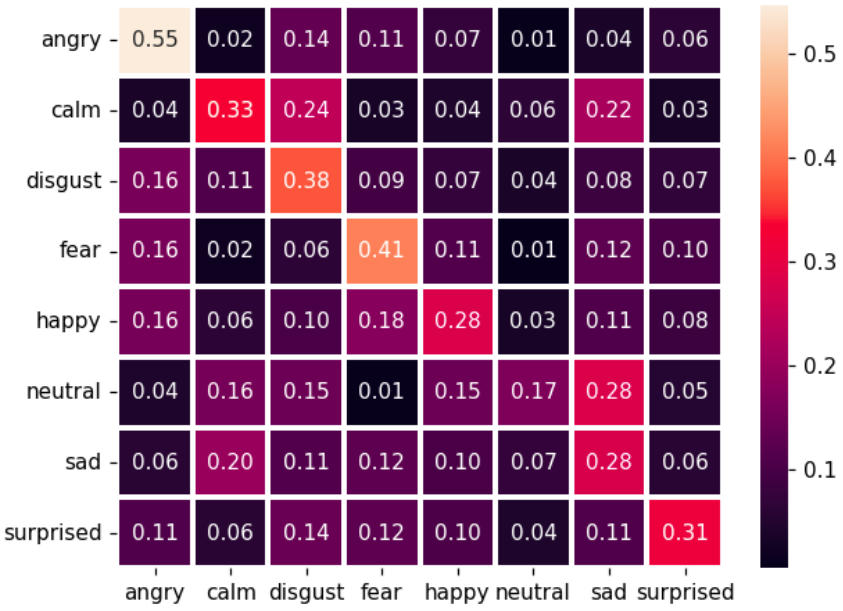
Полнота представляет собой отношение $ИП/(ИП+ЛО)$, где ИП — количество истинных положительных результатов, а ЛО — количество ложноотрицательных результатов. Также под полнотой понимается интуитивно способность классификатора находить все положительные образцы.

Значение полноты лежит в диапозоне от 0 до 1.

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)



Заключение.
TBD

Список литературы

[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.

EXTRACTING MEANINGFUL CHARACTERISTICS OF AN AUDIO SIGNAL FOR SPEECH EMOTION RECOGNITION

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

Department of Electronic Computing Facilities
Faculty of Computer Systems and Networks
Belarusian State University of computer science and Radio Electronics, Republic of Belarus
E-mail: daniil.krasnoproshin@gmail.com

Abstract. TDB
Keywords: DSP, audio feature extraction, recognition, machine learning, data processing.