Распознавание эмоций в речи с помощью глубокого обучения

Мохамед Гисмельбари[1], Е. Е. Гоголев[2], Г. М. Ковалев[3], И. И. Викснин[4]

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[1]mohamedgbari@gmail.com, [2]stavrwalker@gmail.com, [3]grishyc@gmail.com, [4]wixnin@mail.ru

***Аннотация*. В данном исследовании рассматривается применение методов глубокого обучения для распознавания эмоциональных состояний по разговорной речи. В частности, используются сверточные нейронные сети (CNN) и модель HuBERT для анализа аудиовизуальной базы данных эмоциональной речи и песен Райерсона (Ryerson Audio-Visual Database of Emotional Speech and Song, RAVDESS). Полученные результаты свидетельствуют о том, что модели глубокого обучения, в частности модель HuBERT, демонстрируют значительный потенциал в точном определении эмоций в речи. Модели были обучены и протестированы на наборе данных, содержащем различные эмоциональные выражения, включая счастье, печаль, гнев, страх и другие. Эксперименты включали предварительную обработку аудиоданных, извлечение признаков с помощью частотных цепстральных коэффициентов (MFCC) и применение архитектур глубокого обучения для классификации эмоций. Модель HuBERT с усовершенствованным механизмом автоматического обучения превзошла традиционные CNN по точности и эффективности. Это исследование подчеркивает важность выбора подходящих моделей глубокого обучения и наборов признаков для задачи распознавания речевых эмоций. Анализ показывает, что модель HuBERT, использующая контекстную информацию и временную динамику речи, предлагает перспективный подход для разработки более чувствительных и точных систем SER.**

Ключевые слова: Распознавание эмоций в речи; глубокое обучение; сверточные нейронные сети; модель HuBERT; набор данных RAVDESS

I. Введение

Прогресс в области обработки разговорной речи, пересекающийся с обработкой естественного языка, когнитивными науками и человеко-машинным взаимодействием (ЧМС), значительно ускорил развитие адаптивных и отзывчивых человеко-машинных интерфейсов. Распознавание эмоций в речи (Speech Emotion Recognition, SER) становится важнейшим компонентом, повышающим естественность и эффективность человеко-машинных диалоговых систем. Данное исследование посвящено применению методов глубокого обучения для SER, используя базу данных Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) для обучения и оценки моделей глубоких нейронных сетей (DNN). В исследовании проводится обширный обзор литературы, посвященный современным методологиям и достижениям в области SER. В нем предпринята попытка построить полный цикл системы SER, использующий модели глубокого обучения для обнаружения эмоций по аудиоданным. Целью проекта является реализация модели SER, оценка ее эффективности на основе точности, прецизионности, чувствительности и F1-метрики, а также тонкая настройка этой модели для оптимизации производительности. Кроме того, сравнивается эффективность предварительно обученных моделей, таких как HuBERT и ResNet, с моделями, разработанными вручную.

II. Системы распознавания эмоций в речи

Область искусственного интеллекта (ИИ) стала катализатором значительных достижений в области человеко-машинного взаимодействия, в частности, благодаря разработке системы распознавания эмоций речи (SER). SER получила широкую известность благодаря своей способности различать нюансы эмоциональных состояний в человеческой речи, что очень важно для различных приложений, включая развлечения, автомобильную безопасность, виртуальных помощников, здравоохранение, центры обслуживания клиентов и платформы электронного обучения.

Точное распознавание эмоций по речи представляет собой сложную задачу из-за вариативности речи и тонкости эмоциональных выражений. Для достижения высокого уровня производительности SER необходимо пройти через сложные процессы, включая предварительную обработку аудиоданных, извлечение признаков и классификацию эмоций. Усовершенствование возможностей SER продолжает оставаться методологическим поиском алгоритмов, способных превзойти существующие стандарты, что подчеркивает сложность и многомерность человеческих эмоций, передаваемых посредством речи.

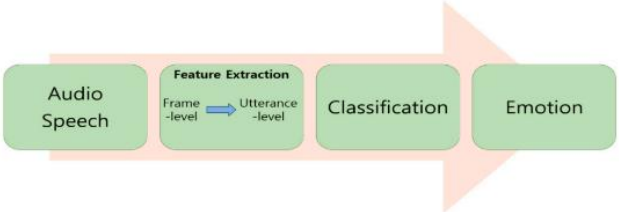


Рис. 1. Процесс работы SER

III. Распространенные аудио признаки, извлекаемые для решения задач классификации звука

В рамках любого проекта машинного обучения, связанного с аудио, начальным этапом является сбор

данных об аудиосигнале, которые затем должны быть преобразованы в признаки, пригодные для алгоритмической обработки. Ключевой частью этого процесса является определение и извлечение наиболее важных аудио характеристик для построения модели, в частности для задач классификации аудио.

A. Частотные цепстральные коэффициенты (MFCC)

Используя библиотеку Librosa в Python, можно извлечь такие важные характеристики, как частотные цепстральные коэффициенты (MFCC). MFCC играют ключевую роль в передаче тембральных и текстурных качеств звука в частотной области, приближенной к характеристикам слуховой системы человека. На эти коэффициенты влияет форма голосового тракта человека, включая язык и зубы, что играет важную роль в точном воспроизведении звуков. MFCC эффективно иллюстрируют нюансы восприятия звуков через шкалу Mel, учитывая то, как люди различают разницу в высоте тона в широком диапазоне частот от 20 Гц до 20 кГц.

$$Mel(f) = 2595\log\left(1 + \frac{f}{100}\right) \quad (1)$$

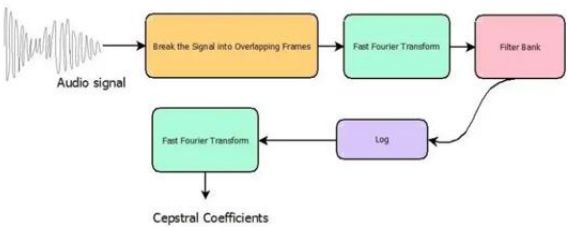


Рис. 2. Блок-схема получения коэффициентов MFCC

B. Оконное преобразование Фурье

Оконное преобразование Фурье (STFT) – это метод, который применяет преобразование Фурье к частям сигнала для получения частотной информации, локализованной во времени. В отличие от стандартного преобразования Фурье, которое дает среднюю частотную оценку для всего сигнала, STFT позволяет уловить, как частотные компоненты изменяются во времени. Это достигается путем разделения сигнала на блоки фиксированного размера (например, 2048 отсчетов) и преобразования каждого из них отдельно. В результате получается спектрограмма, которая отображает время, частоту и амплитуду, обеспечивая полное представление сигнала. Эта спектрограмма служит основой для извлечения звуковых признаков.

Рис. 3. Оконное преобразование Фурье звукового сигнала

C. Насыщенность

Признак насыщенности (Chroma), качество класса высоты тона, относящееся к «цвету» музыкального тона, который может быть разложен на октавно-инвариантное значение, называемое «цветностью», и «высоту тона», указывающую на октаву, в которой находится тон.

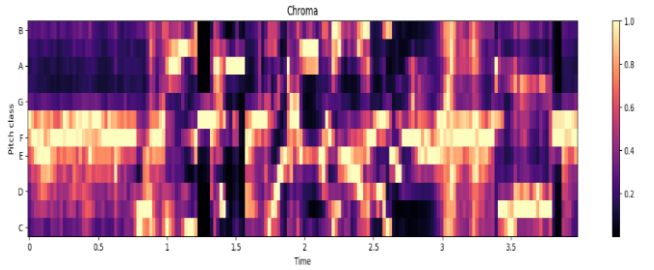


Рис. 4. Типичная хроматическая спектрограмма звукового сигнала

Звуковые свойства, описанные в этом разделе, будут извлечены из аудиосигнала и помещены в массив значений, которые затем будут поданы в модель для классификации эмоций. Это определяет процесс предварительной обработки для аудиоданных, как показано на рис. 5.

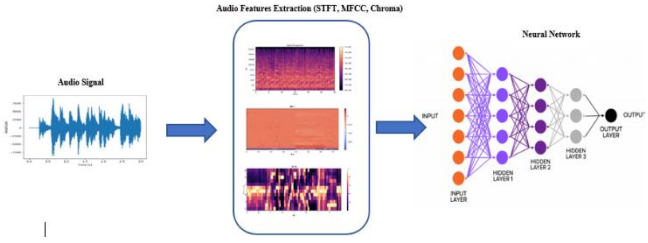


Рис. 5. Процесс предварительной обработки аудиоданных

IV. Набор данных RAVDESS

Аудиовизуальная база данных эмоциональной речи и песен Райерсона (Ryerson Audio-Visual Database of Emotional Speech and Song, RAVDESS) содержит 1440 файлов. RAVDESS содержит 24 профессиональных актера (12 женщин, 12 мужчин), озвучивающих два лексически сопоставимых высказывания с нейтральным североамериканским акцентом. Речевые эмоции включают в себя спокойствие, радость, грусть, гнев, страх, удивление и отвращение. Каждая эмоция производится на двух уровнях эмоциональной интенсивности (нормальный, сильный), а также нейтральном уровне.

V. Подготовка данных

В этом разделе описаны этапы предварительной обработки аудиосигнала перед его подачей в модель глубокого обучения для обучения, тестирования и последующей классификации эмоций из речи. Эти этапы можно определить следующим образом:

- После импортирования всего набора данных в скрипт на вход подается пример аудио из набора данных RAVDESS. Из набора данных извлекаются только четыре эмоции: нейтральная, грустная, сердитая и счастливая.
- Был создан цикл по каталогу RAVDESS для сбора из аудиопапок эмоций, пола говорящего и построена диаграмма всех собранных аудиофайлов и эмоций.
- Каждый из этих аудиофайлов попадает в процесс извлечения признаков, как описано в разделе III.

- Все извлеченные признаки данных эмоций объединяются в фрейм данных, который будет служить входными данными для CNN-модели для классификации эмоций.

```
neutral    192
happy      192
sad        192
angry      192
Name: emotion, dtype: int64
```

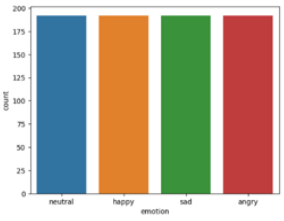


Рис. 6. Количество предполагаемых эмоций

VI. Обучение и тестирование CNN модели

Традиционная архитектура CNN, описанная в литературе, предназначена для классификации эмоций. Эта модель включает в себя восемь сверточных слоев, использующих функцию активации ReLU. Для снижения риска переобучения в модель включены отсеивающие слои. Архитектура начинается с первого сверточного слоя, который инициализируется в соответствии с размерами входной переменной x_train, а именно (218,1). Завершает модель полносвязный слой с четырьмя элементами, отражающими количество целевых классов эмоций. Для облегчения процесса обучения в модели используется оптимизатор Adam с заданной скоростью обучения 0,0001. Тестовая выборка составляет 20 % наблюдений, а тренировочная – 80 %.

Точность на тестовом наборе данных составила 40 %. В качестве функции потерь была выбрана «категориальная перекрестная энтропия». Данная точность оказалась максимальной, достигнутой данной моделью, даже после настройки параметров. Поэтому было решено использовать другую модель для достижения более высокой точности при тех же исходных данных. В качестве такой модели была выбрана модель HuBERT, которая будет рассмотрена в следующем разделе.

VII. Обучение и тестирование модели HuBERT

Модель HuBERT, построенная на архитектуре «Трансформер», специально разработана для таких задач, как распознавание эмоций в речи. Она однозначно определяет дискретные элементы в речи, что позволяет глубже понять эмоциональные сигналы. Обучение HuBERT начинается с предварительного обучения без учителя. Модель обучается на неразмеченных данных, предсказывая маскированные речевые сегменты и развивая общее понимание речевых шаблонов. Затем она проходит тонкую настройку на основе маркированных данных из RAVDESS, фокусируясь на четырех указанных эмоциях. Затем проводится тестирование модели на новых данных, оценивается ее точность при идентификации эмоций. Точность модели HuBERT составила 83,77 % по сравнению с 40 % у модели CNN.

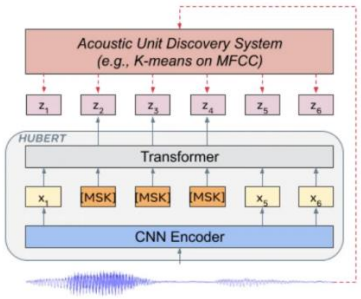


Рис. 7. Архитектура модели HuBERT

VIII. Анализ результатов исследования

В этой секции подробно рассматриваются результаты работы двух моделей распознавания эмоций в речи (SER): пользовательской CNN-модели и базовой модели HuBERT, использующих набор данных RAVDESS. Базовая модель HuBERT значительно превзошла модель CNN, достигнув общей точности 83,77 % по сравнению с 40,97 % у CNN. В этом разделе будет проведено подробное сравнение моделей, показаны их сильные и слабые стороны в SER с помощью матриц ошибок и ключевых оценочных показателей, таких как точность, чувствительность и F1 метрика.

A. Матрица ошибок

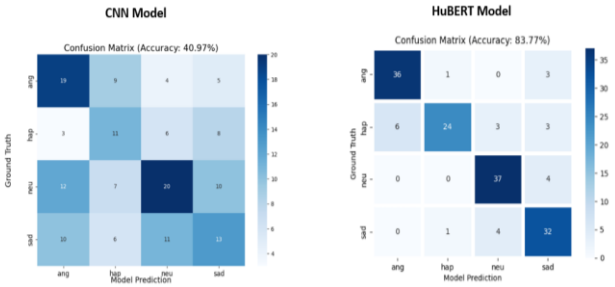


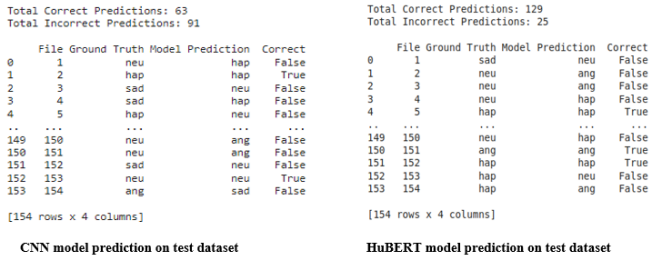
Рис. 8. Матрицы ошибок CNN и модели HuBERT

Изучение матриц ошибок, представленных на рис. 8, показывает явное различие в производительности двух моделей, особенно заметное в диагональных сегментах, которые обозначают точные классификации эмоций. Использование цветных диаграмм, на которых более темные и более светлые оттенки синего означают более высокую и более низкую точность соответственно, визуально указывает на превосходство модели HuBERT в распознавании эмоций. Для более полного анализа были использованы дополнительные метрики оценки, рассмотренные в следующем разделе.

B. Предсказания моделей на тестовом наборе данных

Точность модели на тестовом наборе данных определяется следующим образом:

$$\frac{\text{Общее количество правильных предсказаний}}{\text{Общее количество предсказаний}} * 100 \qquad (2)$$



Total Correct Predictions: 63
Total Incorrect Predictions: 91

	File	Ground Truth	Model Prediction	Correct
0	1	neu	hap	False
1	2	hap	hap	True
2	3	sad	neu	False
3	4	sad	hap	False
4	5	hap	neu	False
..	...	...	...	...
149	150	neu	ang	False
150	151	neu	ang	False
151	152	sad	neu	False
152	153	neu	neu	True
153	154	ang	sad	False

[154 rows x 4 columns]

CNN model prediction on test dataset

Total Correct Predictions: 129
Total Incorrect Predictions: 25

	File	Ground Truth	Model Prediction	Correct
0	1	sad	neu	False
1	2	neu	ang	False
2	3	neu	ang	False
3	4	neu	hap	False
4	5	hap	hap	True
..	...	...	...	...
149	150	neu	hap	False
150	151	ang	ang	True
151	152	hap	hap	True
152	153	hap	neu	False
153	154	hap	ang	False

[154 rows x 4 columns]

HuBERT model prediction on test dataset

Рис. 9. Точность предсказания моделей CNN и HuBERT на тестовом наборе данных

На рис. 9 показано, что модель HuBERT значительно превосходит модель CNN по точности на тестовом наборе данных, достигнув 129 правильных предсказаний из 154 возможных. В отличие от этого, модель CNN смогла сделать только 63 правильных прогноза, что подчеркивает превосходство модели HuBERT.

Точность модели CNN на тестовом наборе данных:

$$\frac{63}{154} * 100 = 40.97\% \quad (3)$$

Точность модели HuBERT на тестовом наборе данных:

$$\frac{129}{154} * 100 = 83.76\% \quad (4)$$

C. Точность, чувствительность, F1-мера

	precision	recall	f1-score
ang	0.25	0.27	0.26
hap	0.20	0.20	0.20
neu	0.22	0.20	0.21
sad	0.28	0.30	0.29

CNN model evaluation metrics

Emotion	Precision	Recall	F1-Score
ang	0.230769	0.300000	0.260870
hap	0.296296	0.355556	0.323232
neu	0.428571	0.394737	0.410959
sad	0.461538	0.292683	0.358209

HuBERT model evaluation metrics

Рис. 10. Метрики оценки 4 эмоций для CNN и модели HuBERT

На рис. 10 показано, что модель HuBERT значительно превосходит модель CNN, получая более высокие оценки по всем метрикам, особенно заметно значительное увеличение точности и F1-меры для эмоций «нейтральный» и «грустный» по сравнению с моделью CNN.

IX. Заключение

В данном исследовании успешно реализованы и сравнены две модели распознавания эмоций в речи (SER) – пользовательская CNN и базовая модель HuBERT – на наборе данных RAVDESS, сфокусированном на четырех основных эмоциях: нейтральной, грустной, злой и счастливой.

Основные результаты:

- Разработана комплексная система SER, включающая в себя сбор аудиоданных и распознавание эмоций.
- Изучение распространенных моделей нейронных сетей (CNN, LSTM, RNN) и создание собственной модели CNN.
- Изучение и использование предварительно обученных моделей, таких как HuBERT и ResNet, причем HuBERT была выбрана за ее превосходную производительность.
- Тонкая настройка обеих моделей для повышения производительности, подчеркивающая важность оптимизации гиперпараметров.

Анализ производительности показал, что модель HuBERT значительно превосходит модель CNN, достигая точности 83,77 % против 40,97 %, что делает ее предпочтительным выбором для оптимальной работы SER.

Список литературы

[1] *Sanjita. B. R, Nipunika. A, Rohita Desai,* Speech Emotion Recognition using MLP Classifier

[2] *Harini Murugan.* Speech Emotion Recognition Using CNN

[3] *Florian Eyben et al.* The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing // IEEE

[4] George Trigeorgis et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network // ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing. 2016. С. 5200-5204.

[5] Transactions on Affective Computing 7.2. 2016. С. 190-202.

[6] *Jianfeng Zhao, Xia Mao, and Lijiang Chen*. Speech emotion recognition using deep 1D & 2D CNN LSTM networks // Biomedical Signal Processing and Control 47. 2019. С. 312-323.

[7] *Klaus R. Scherer*. Vocal communication of emotion: A review of research paradigms // Speech Communication 40.1-2. 2003 С. 227-256.

[8] *S.R. Livingstone and F.A. Russo*. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). 2018. С. 1–35.

[9] *Siqing Wu, Tiago H. Falk, and Wai Yip Chan*. Automatic speech emotion recognition using modulation spectral features // Speech Communication 53.5. 2011. С. 768–785.

[10] *Mittal, A., Arora, V., & Kaur, H*. Speech Emotion Recognition using HuBERT Features and Convolutional Neural Networks. // 2021 6th International Conference on Computing, Communication and Security (ICCCS) IEEE. 2021. С.1-8.

[11] *Zhang, Y., Yang, Y., Li, Y., Li, W., & Zhao, J.* Speech emotion recognition based on HuBERT and attention mechanism. // Proceedings of the 2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE) IEEE. С. 278-280.