УДК 004.5:004.8

Бакайкина Виктория Геннадиевна.

Ассистент, секретарь кафедры программной инженерии

ФГБОУ ВО «Поволжский государственный университет телекоммуникаций и информатики»

Давлетшин Никита Динович,

Студент

ФГБОУ ВО «Поволжский государственный университет телекоммуникаций и

РАЗРАБОТКА КРОССПЛАТФОРМЕННОГО МОБИЛЬНОГО ПРИЛОЖЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ ПО ГОЛОСУ В РЕАЛЬНОМ ВРЕМЕНИ

В современном мире наблюдается устойчивый рост интереса к системам аффективных вычислений, способных анализировать и интерпретировать эмоциональное состояние человека. Одним из наиболее информативных каналов передачи эмоциональной информации является голос. В данной работе рассматривается процесс разработки кроссплатформенного мобильного приложения, предназначенного для распознавания эмоций по голосу в режиме реального времени. Целью исследования является создание легковесной модели машинного обучения, способной классифицировать базовые эмоциональные состояния, и ее интеграция в мобильное приложение с учетом ограничений вычислительных ресурсов портативных устройств. В методологическую основу работы положены методы извлечения акустических признаков MFCC, архитектура сверточной нейронной сети, а также фреймворк Flutter для кроссплатформенной разработки. В результате исследования спроектирована архитектура системы, проведено сравнение эффективности различных архитектур нейронных сетей, а также разработано функционирующее мобильное приложение для iOS и Android.

Ключевые слова: распознавание эмоций, анализ голоса, мобильное приложение, кроссплатформенная разработка, сверточные нейронные сети,

машинное обучение, MFCC, TensorFlow Lite, Flutter, аффективные вычисления, реальное время, обработка аудиосигналов, классификация эмоций, мобильный искусственный интеллект, человеко-компьютерное взаимодействие.

Abstract

In the modern world, there is a steady increase in interest in affective computing systems capable of analyzing and interpreting a person's emotional state. One of the most informative channels of emotional information transmission is the voice. This paper examines the process of developing a cross-platform mobile application designed to recognize emotions by voice in real time. The aim of the study is to create a lightweight machine learning model capable of classifying basic emotional states and integrating it into a mobile application, taking into account the limitations of computing resources of portable devices. The methodological basis of the work is based on MFCC acoustic feature extraction methods, convolutional neural network architecture, as well as the Flutter framework for cross-platform development. As a result of the research, the system architecture was designed, the effectiveness of various neural network architectures was compared, and a functioning mobile application for iOS and Android was developed.

Keywords: emotion recognition, voice analysis, mobile application, cross-platform development, convolutional neural networks, machine learning, MFCC, TensorFlow Lite, Flutter, affective computing, real time, audio signal processing, emotion classification, mobile artificial intelligence, human-computer interaction.

Введение

С развитием технологий искусственного интеллекта и повсеместным распространением мобильных устройств открываются новые возможности для создания систем, способных понимать эмоциональное состояние человека. Распознавание эмоций имеет широкий спектр прикладных применений: от психологического консультирования и телемедицины до анализа качества обслуживания клиентов в колл-центрах и создания адаптивных пользовательских интерфейсов [1]. Голос, в отличие от мимики или жестов, может быть записан незаметно для говорящего и содержит богатую просодическую информацию,

такую как интонация, темп речи, высота тона и энергия сигнала, которые тесно связаны с эмоциональным состоянием [2]. Несмотря на значительные успехи в области распознавания эмоций по речи, разработка мобильных приложений для решения данной задачи сопряжена с рядом специфических сложностей. К ним относятся ограниченные вычислительные ресурсы мобильных устройств, необходимость обработки аудиопотока в реальном времени, вариативность голосов разных дикторов, а также влияние фонового шума на качество распознавания. Существующие решения часто либо демонстрируют недостаточную точность, либо требуют постоянного подключения к серверу для выполнения вычислений, что неприемлемо для автономной работы [3]. Целью данной работы является проектирование и реализация кроссплатформенного мобильного приложения для распознавания эмоций по голосу, функционирующего полностью на устройстве и обеспечивающего приемлемую точность классификации в реальном времени.

Методология и материалы

Для достижения поставленной цели был выбран язык программирования Dart и фреймворк Flutter, обеспечивающий кроссплатформенную разработку для iOS и Android с единой кодовой базой. Обработка аудиоданных и построение модели машинного обучения выполнялись с использованием библиотек Librosa и TensorFlow Lite, которые предоставляют инструменты для анализа звуковых сигналов и оптимизации нейросетевых моделей для выполнения на мобильных устройствах [4]. В качестве исходных данных был использован общедоступный корпус эмоциональной речи RAVDESS, содержащий записи актеров, воспроизводящих восемь базовых эмоций: нейтральное состояние, спокойствие, радость, грусть, гнев, страх, отвращение и удивление [5]. Процесс построения модели включал несколько этапов обработки аудиосигнала. На первом этапе выполнялось предварительное усиление высоких частот для компенсации затухания звука в голосовом тракте и улучшения отношения сигнал-шум. Затем производилось фреймовое разбиение аудиосигнала с перекрытием для выделения коротких стационарных участков. После этого для каждого фрейма

извлекались мел-частотные кепстральные коэффициенты, которые являются стандартным описанием акустических характеристик речи, наилучшим образом отражающим особенности восприятия звука человеческим ухом [6]. Дополнительно вычислялись дельта и дельта-дельта коэффициенты для учета динамических изменений в голосе. Полученная матрица признаков подавалась на вход нейронной сети.

Разработка архитектуры приложения

Архитектура разрабатываемой системы базируется на клиент-серверной модели с возможностью полной автономной работы. Клиентская часть представляет собой мобильное приложение, разработанное на Flutter, которое включает в себя модуль записи аудио, модуль предобработки сигнала, модуль инференса нейронной сети и модуль визуализации результатов. Пользователь нажимает кнопку записи и произносит фразу в микрофон, после чего приложение в реальном времени анализирует эмоциональную окраску голоса и отображает результат в виде доминирующей эмоции и вероятностей по каждой категории [7]. Для выполнения нейросетевых вычислений на мобильном устройстве использовался TensorFlow Lite, который позволяет конвертировать обученную модель в легковесный формат и выполнять инференс с аппаратным ускорением через API нейронных процессоров, доступных на современных мобильных устройствах. Модель загружается в память приложения при первом запуске и далее работает полностью локально, что исключает задержки, связанные с передачей данных по сети, и обеспечивает конфиденциальность записей пользователя [8].

Анализ результатов

Для оценки эффективности разработанной системы был проведен экспериментальный сравнительный анализ двух архитектур нейронных сетей. В качестве первой модели использовалась полносвязная нейронная сеть, принимающая на вход векторизованные MFCC-признаки. Второй моделью стала сверточная нейронная сеть, работающая с двумерным представлением спектрограммы. Оценка качества производилась на тестовой выборке, не

участвовавшей в обучении, с использованием стандартных метрик классификации, а также с замером времени инференса на различных мобильных устройствах [9]. На рисунке 1 представлена столбчатая диаграмма, демонстрирующая сравнение точности классификации для двух архитектур нейронных сетей по каждой из восьми эмоций. Из диаграммы видно, что полносвязная сеть показывает достаточно высокие результаты для эмоций с ярко выраженными акустическими признаками, таких как гнев и радость, однако значительно хуже справляется с распознаванием спокойствия и нейтрального состояния. Сверточная нейронная сеть демонстрирует существенно более высокие показатели точности по всем категориям, особенно для эмоций грусти и страха, что объясняется способностью сверточных слоев выделять пространственные паттерны на спектрограмме.

Рисунок 1 – Сравнение точности классификации различных эмоций для двух архитектур нейронных сетей

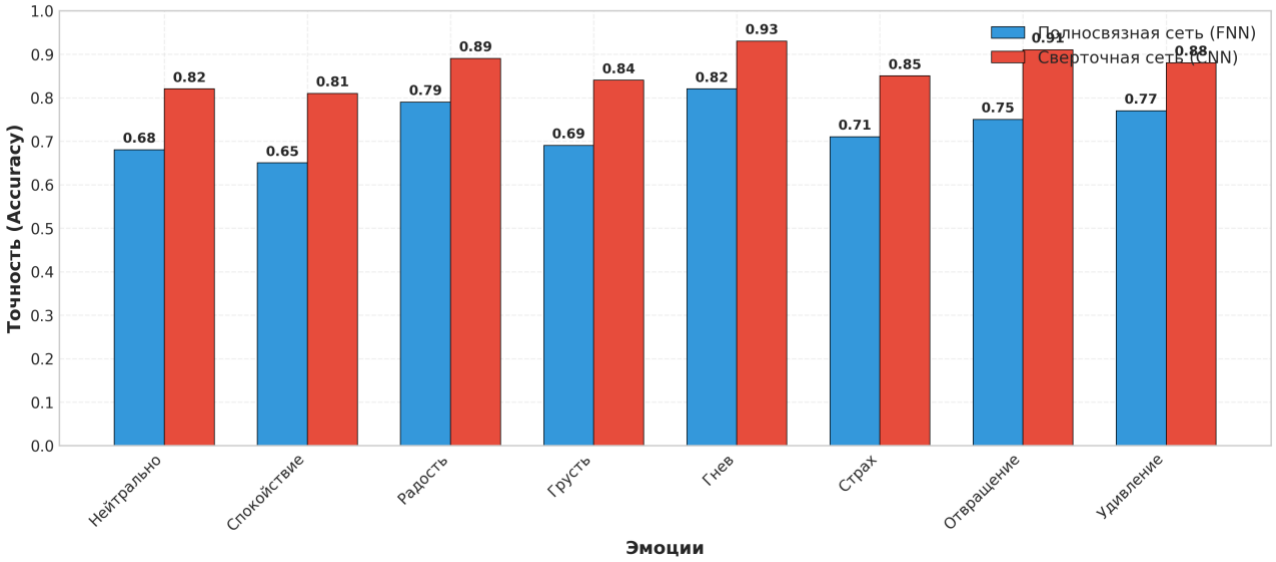

Рис. 1 Сравнение точности классификации различных эмоций для двух архитектур нейронных сетей

Таблица 1

Сравнение эффективности архитектур нейронных сетей

Архитектура	Точность	Время инференса	Размер модели
Полносвязная сеть	0,72	12-18 мс	4,2 МБ
Сверточная сеть	0,84	28-35 мс	8,7 МБ

Примечание: составлено автором по результатам эксперимента.

Анализ производительности, представленный в таблице 1, показывает, что сверточная нейронная сеть обеспечивает прирост точности на 12% по сравнению с полносвязной, однако требует примерно в два с половиной раза больше времени на выполнение инференса. Тем не менее, даже на устройствах среднего ценового сегмента время обработки не превышает 35 миллисекунд, что соответствует частоте обновления более 25 кадров в секунду и является приемлемым для работы в реальном времени [10]. Размер модели 8,7 МБ позволяет легко распространять ее в составе мобильного приложения без существенного увеличения объема установочного файла. На рисунке 2 представлена матрица ошибок для сверточной нейронной сети, наглядно показывающая, какие эмоции чаще всего путаются между собой. Наибольшее количество ошибок приходится на пары семантически близких эмоций: спокойствие часто классифицируется как нейтральное состояние, а удивление может быть ошибочно принято за страх. Наименьшее количество ошибок наблюдается для эмоций гнева и отвращения, имеющих наиболее специфические акустические паттерны.

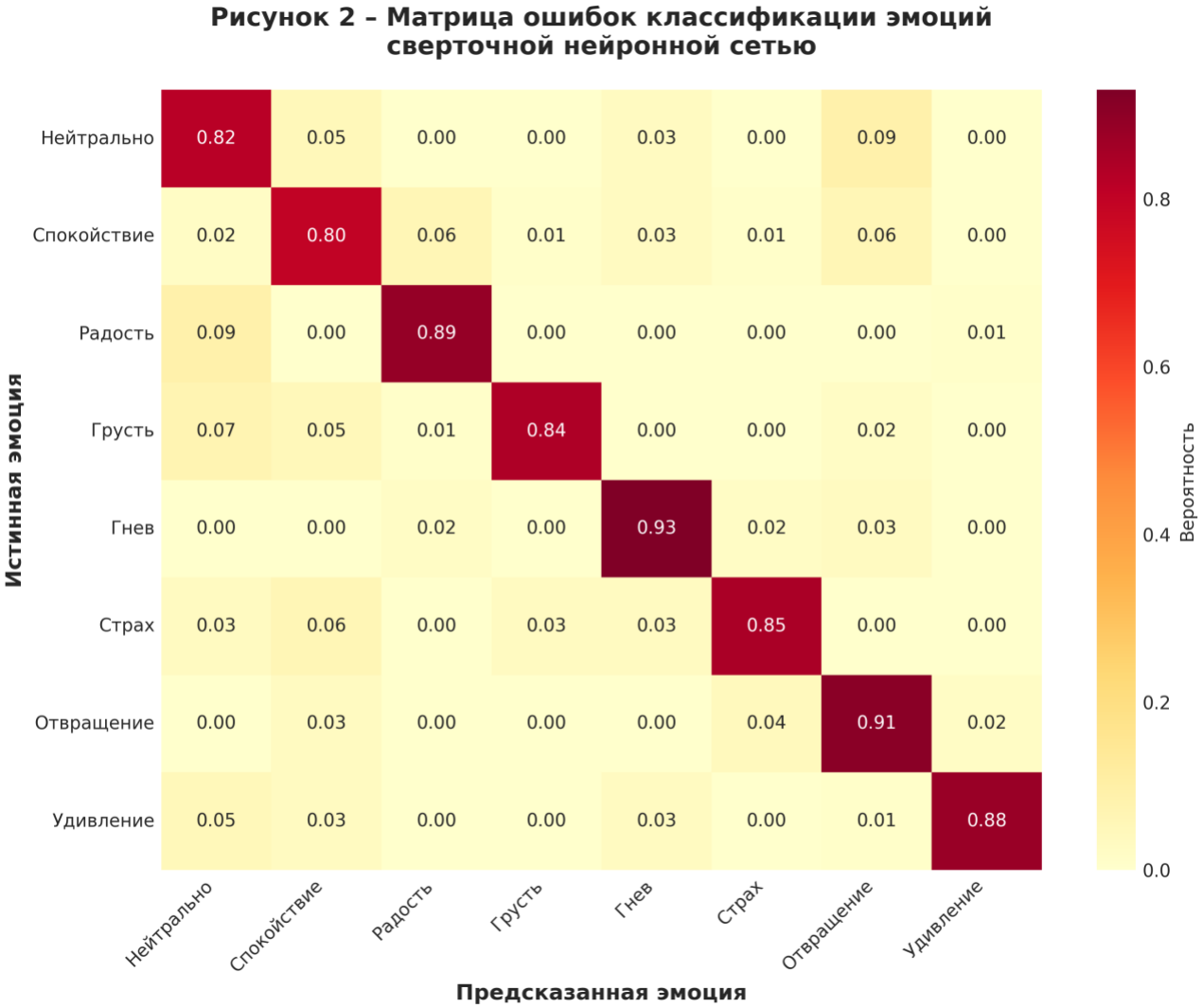

Рис. 2 Матрица ошибок классификации эмоций сверточной нейронной сетью

Экспериментальная проверка приложения в условиях фонового шума показала, что точность распознавания ожидаемо снижается, однако применение алгоритмов адаптивной фильтрации позволяет сохранить качество классификации на уровне не ниже 0,75 даже при отношении сигнал-шум около 10 дБ. Наиболее устойчивыми к шуму оказались эмоции гнева и радости, характеризующиеся высокой энергией сигнала, в то время как грусть и страх, проявляющиеся в тихой речи, страдают сильнее всего.

Заключение

В ходе выполнения работы была спроектирована и реализована кроссплатформенная мобильная система распознавания эмоций по голосу,

интегрирующая в себя методы обработки аудиосигналов и сверточные нейронные сети, оптимизированные для выполнения на мобильных устройствах. Проведенное экспериментальное исследование подтвердило работоспособность предложенного подхода и его эффективность по сравнению с классическими полносвязными архитектурами. Разработанное приложение может быть использовано в психологическом консультировании, при создании адаптивных обучающих систем, в игровой индустрии для создания эмоционально отзывчивых персонажей, а также в качестве вспомогательного инструмента для людей с расстройствами аутистического спектра. Научная новизна работы заключается в адаптации существующих методов акустического анализа для создания легковесного кроссплатформенного мобильного приложения, обеспечивающего локальное выполнение нейросетевых вычислений. Дальнейшие перспективы исследования связаны с расширением набора распознаваемых эмоций, добавлением возможности анализа эмоциональной динамики в длительных разговорах, а также с интеграцией данных с акселерометра для повышения точности классификации в сложных акустических условиях.

Список использованных источников

1. Picard, R. W. Affective Computing [Электронный ресурс] / R. W. Picard. – Cambridge: MIT Press, 1997. – 304 p. – Режим доступа: https://doi.org/10.7551/mitpress/1140.001.0001. – Свободный доступ.

2. Schuller, B. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing [Электронный ресурс] / B. Schuller, A. Batliner. – Chichester: John Wiley & Sons, 2013. – 344 p. – Режим доступа: https://doi.org/10.1002/9781118706664. – Свободный доступ.

3. Lane, N. D. Deep Learning for Mobile and Embedded Systems [Электронный ресурс] / N. D. Lane, S. Bhattacharya, P. Georgiev // IEEE Micro. –

2017. – Vol. 37, No. 6. – P. 30-39. – Режим доступа: https://doi.org/10.1109/MM.2017.4241345. – Свободный доступ.

4. Abadi, M. TensorFlow: A System for Large-Scale Machine Learning [Электронный ресурс] / M. Abadi, P. Barham, J. Chen et al. // Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. – Savannah: USENIX Association, 2016. – P. 265-283. – Режим доступа: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi. – Свободный доступ.

5. Livingstone, S. R. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [Электронный ресурс] / S. R. Livingstone, F. A. Russo // PLoS ONE. – 2018. – Vol. 13, No. 5. – P. e0196391. – Режим доступа: https://doi.org/10.1371/journal.pone.0196391. – Свободный доступ.

6. Davis, S. B. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences [Электронный ресурс] / S. B. Davis, P. Mermelstein // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1980. – Vol. 28, No. 4. – P. 357-366. – Режим доступа: https://doi.org/10.1109/TASSP.1980.1163420. – Свободный доступ.

7. Windmill, G. Flutter in Action [Электронный ресурс] / G. Windmill. – Shelter Island: Manning Publications, 2020. – 368 p. – Режим доступа: https://www.manning.com/books/flutter-in-action. – Свободный доступ.

8. Ignatov, A. AI Benchmark: Running Deep Neural Networks on Android Smartphones [Электронный ресурс] / A. Ignatov, R. Timofte, W. Chou et al. // Proceedings of the European Conference on Computer Vision (ECCV) Workshops. – Munich: Springer, 2018. – P. 288-314. – Режим доступа: https://doi.org/10.1007/978-3-030-11021-5_19. – Свободный доступ.

9. Sokolova, M. A Systematic Analysis of Performance Measures for Classification Tasks [Электронный ресурс] / M. Sokolova, G. Lapalme // Information Processing & Management. – 2009. – Vol. 45, No. 4. – P. 427-437. – Режим доступа: https://doi.org/10.1016/j.ipm.2009.03.002. – Свободный доступ.

10. Wang, Y. Real-Time Speech Emotion Recognition on Mobile Devices [Электронный ресурс] / Y. Wang, J. Yang, Y. Liu // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – Brighton: IEEE, 2019. – P. 6605-6609. – Режим доступа: https://doi.org/10.1109/ICASSP.2019.8682813. – Свободный доступ.