ПРИМЕНЕНИЕ YAMNET В КЛАССИФИКАЦИИ ДЕТСКИХ И ВЗРОСЛЫХ ГОЛОСОВ

А.А. Минсафин, *магистрант*
И.М. Куфтерин, *магистрант*
Л.М. Тыщенко, *магистрант*
Научный руководитель: *А.С. Тощев, канд. техн. наук, доцент*
Казанский федеральный университет
(Россия, г. Казань)

DOI:10.24412/2500-1000-2026-5-1-352-360

Аннотация. *В статье представлено экспериментальное сравнение двух подходов к задаче автоматического определения детского голоса: классической сверточной нейронной сети (CNN), и модели YAMNet, использующей предобученные аудио-эмбеддинги. Для обучения и тестирования обоих методов был сформирован собственный датасет на основе выборки «common_voice_17_0» (русский язык) с учётом двух ключевых факторов – пола (мужской/женский) и возрастной категории (дети/взрослые). Для YAMNet данные предварительно обрабатывались с нормализацией, усечением или дополнением до длительности 3 секунд. В подходе на основе YAMNet аудиосигнал преобразуется в 1024-мерные векторные эмбеддинги, среднее по времени которых используется в качестве входа для полносвязных слоев многослойного перцептрона. В архитектуре CNN сигнал преобразуется в спектрограмму мел-коэффициентов и обрабатывается многослойной сверточной сетью, затем итоговые признаки поступают на полносвязные слои.*

Ключевые слова: *распознавание голоса; детский голос; классификация по возрасту; сверточные нейронные сети; CNN; YAMNet.*

Распознавание речи и анализ голоса занимают значительное место в исследованиях по искусственному интеллекту и машинному обучению. Эти технологии находят применение в различных областях, включая голосовых помощников, системы безопасности и фильтрации контента. Одной из актуальных задач является определение возраста говорящего по его голосу, что имеет важное значение для разграничения информации в интернете и обеспечения безопасности детей. С развитием технологий и стремительным увеличением объема как допустимого, так и недопустимого контента в цифровом пространстве возрастает необходимость в автоматических системах, способных точно определять, говорит ли взрослый или ребенок. Такие системы могут предотвращать несанкционированный доступ детей к неприемлемому контенту, обеспечивая более безопасную среду для их взаимодействия с онлайн-ресурсами [1, 2].

Определение детского голоса представляет собой уникальную задачу из-за ряда специфических характеристик. Детская речь отличается более высокой фундаментальной частотой, большей изменчивостью и наличием особенностей произношения, связанных с развитием речевого аппарата [3]. Эти акустические особенности делают детский голос более сложным для анализа по сравнению с голосом взрослого, что требует разработки специализированных методов и алгоритмов. Исторически, большинство исследований в области распознавания речи фокусировалось на взрослых голосах, что обусловлено доступностью данных и меньшей вариативностью речевых паттернов [4], однако современные подходы на основе нейронных сетей [5, 6] позволяют преодолеть эти ограничения.

С внедрением нейронных сетей и глубокого обучения стало возможным значительно улучшить качество распознавания речи и классификации голосов. Эти модели обладают способностью обрабатывать сложные и многомерные данные, что критически важно для задач, связанных с анализом акустических характеристик голоса.

В данной статье мы исследуем, насколько эффективно современные модели машинного обучения могут различать голоса взрослых и детей, учитывая их специфические акустиче-

ские характеристики. Сначала будет рассмотрен выбор данных для обучения моделей, где необходимо правильно разбить количество записей по соотношению возраста. Затем будет произведено сравнение архитектур нейронных сетей YAMNet и CNN и полученных на их основе результатов. Основное внимание здесь уделяется разработке и тестированию моделей, которые могут быть использованы в дальнейшем для автоматической фильтрации контента и обеспечения безопасности детей в цифровом пространстве.

Набор данных

В рамках данного исследования для формирования корпуса данных была использована русскоязычная ветка Mozilla Common Voice 17.0 [7].

Из общего массива записей были извлечены первые пять тысяч аудиофайлов вместе с аннотациями пола («male_masculine» и «female_feminine») и возраста («teens», «twenties», «thirties», «fourties») (рис. 1).

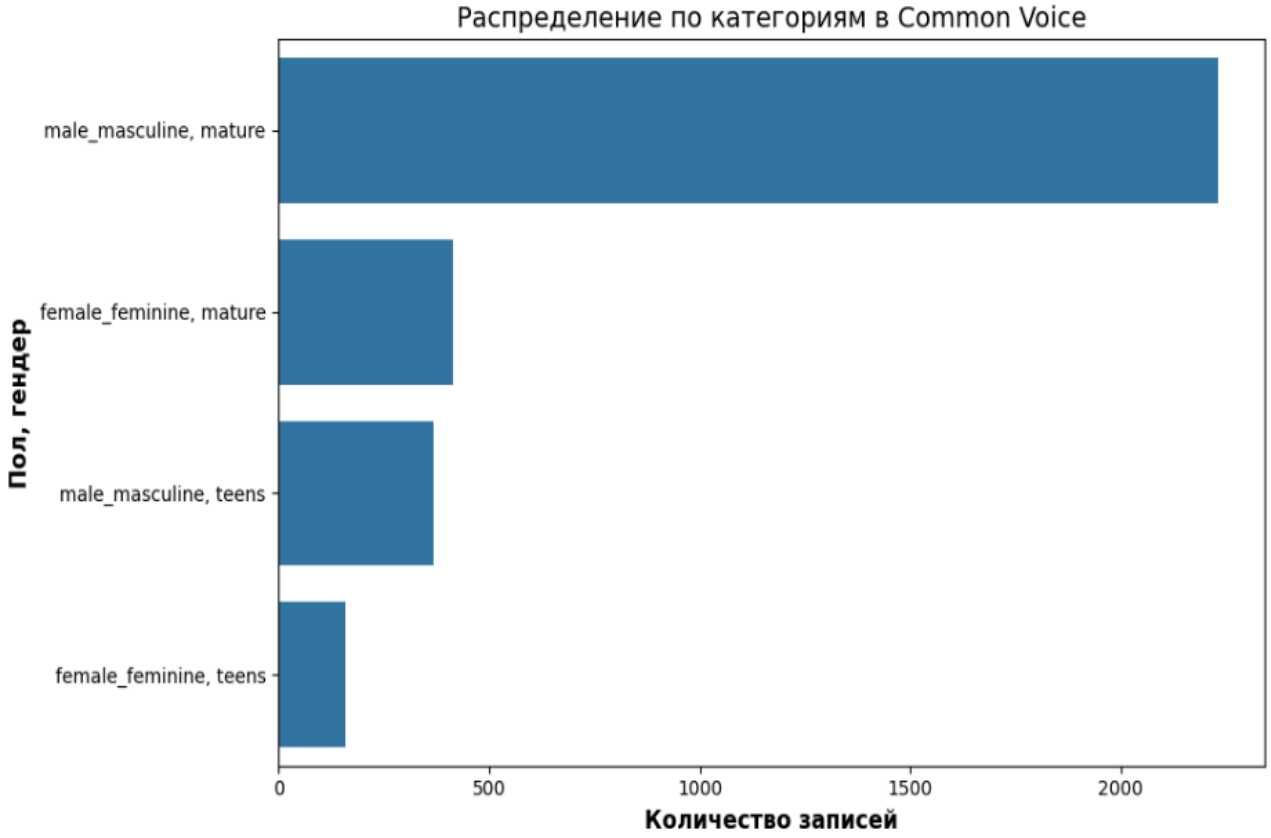

Рис. 1. Распределение по категориям в Mozilla Common Voice

Каждый файл приводился к единому формату: аудиосигнал сначала нормализовался по уровню громкости, затем при необходимости обрезался или дополнялся нулями до фиксированной длины в 3 секунды. После этого весь набор данных разделялся на обучающую и тестовую выборки в пропорции 70% к 30% с сохранением распределения по возрастным категориям для каждого пола.

Для компенсации дисбаланса «детской» категории в обеих выборках применялась аугментация питч-шифтом: к аудиофрагментам добавлялись случайные полутона в диапазоне 1-3 до достижения не менее 100 «детских» примеров в обучении и 50 в тесте. Такой подход позволил выровнять количество экземпляров целевого класса без привлечения дополнительных внешних данных.

Последовательность подготовки данных для сверточной нейронной сети (CNN) включала чтение MP3-файлов средствами TensorFlow I/O, ресемплинг с 48 000 Гц до 16 000 Гц и последующее преобразование в спектрограмму. Полученные амплитудные спектрограммы расширялись дополнительной размерностью канала, что обеспечивало совместимость входов с архитектурой сверточного классификатора.

В случае применения YAMNet после общей нормализации и выравнивания длины аудио прошло через предобученную модель для получения эмбеддингов. Из выходных временных эмбеддингов размерностью 1024 для каждого фрагмента вычислялось среднее по временной оси, и именно этот усреднённый вектор признаков будет использован для обучения классификатора.

В результате предобработки мы получили тренировочные и тестовые выборки для мужского и женского голоса (табл. 1 и 2).

Таблица 1. Распределение для мужских голосов

Выборка	Класс	Количество примеров
Тренировочная	Взрослый голос	1561
	Детский голос	258
Тестовая	Взрослый голос	669
	Детский голос	111

Таблица 2. Распределение для женских голосов

Выборка	Класс	Количество примеров
Тренировочная	Взрослый голос	290
	Детский голос	111
Тестовая	Взрослый голос	125
	Детский голос	50

Архитектура нейросетей

В нашем исследовании для задачи бинарной классификации «детский/взрослый» голоса были выбраны два принципиально различных подхода: с нуля обучаемая сверточная нейронная сеть (CNN) (рис. 2) над спектрограммами и классификатор на основе предобученных эмбеддингов YAMNet (рис. 3). Такой выбор обусловлен необходимостью сравнить прямое извлечение спектро-временных признаков и стратегию transfer learning, актуальную при ограниченных объёмах данных.

Для CNN преобразование аудиофайлов в мел-спектрограммы и последовательное применение сверточных позволяет модели «видеть» локальные гармоники, форманты и переходы, характерные для детского голоса [5, 8]. Глубина трёх сверточных блоков обеспечивает постепенное усложнение признаков – от низкоуровневых тоновых характеристик до более абстрактных структур речи – а полносвязные слои с регуляризацией через Dropout способствуют обобщению, снижая риск переобучения на ограниченном объёме выборки.

Использование YAMNet с передовым подходом transfer learning обосновано необходимостью получать устойчивые аудио-представления при относительно небольшом наборе размеченных данных [2, 6]. Модель, предобученная на обширном аудиокорпусе AudioSet, уже умеет выделять универсальные эмбеддинги, описывающие широкий спектр звуковых событий. Усреднение 1024-мерных эмбеддингов по временной оси превращает переменную длину входного сигнала в компактный вектор признаков, который затем легко адаптируется к задаче классификации детского голоса с помощью полносвязного классификатора. Такой подход позволяет существенно сократить время дообучения и повысить устойчивость результатов в условиях дефицита данных.

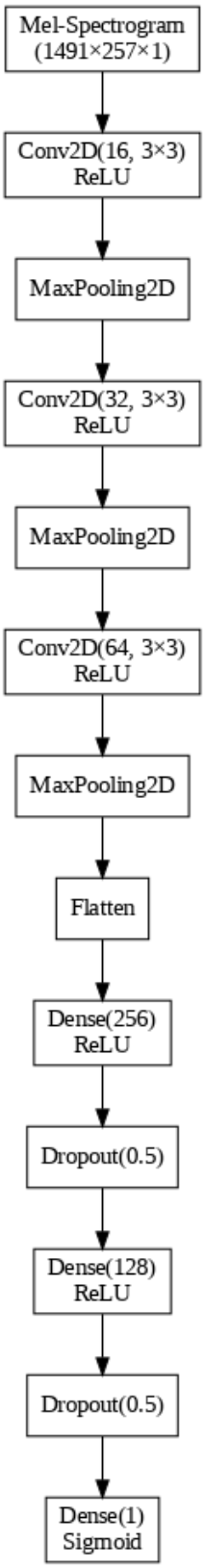

Рис. 2. Архитектура классификатора CNN

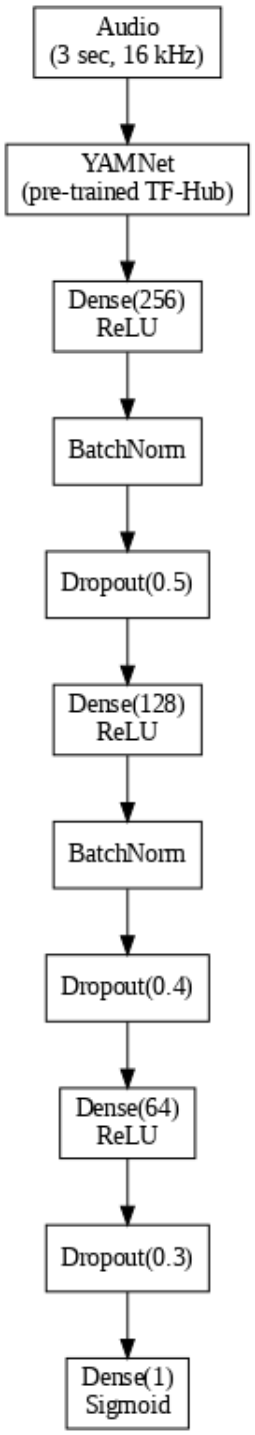

Рис. 3. Архитектура классификатора на основе YAMNet

Обучение классификатора на основе CNN

Процесс обучения CNN проводился в рамках 10 эпох. В качестве оптимизатора для сверточной сети был выбран алгоритм Adam. В качестве функции потерь была выбрана бинарная кросс-энтропия, которая напрямую соответствует задаче двухклассовой классификации «детский/взрослый». Для оценки качества обучения одновременно контролиро-

вались метрики полноты (recall) и точности (precision), что позволило сбалансированно отследить как способность модели находить все примеры целевого класса, так и долю корректных положительных предсказаний в условиях неравномерного распределения классов. В качестве результатов представлены график функции потерь и графики метрик recall и precision (рис. 4-6 и табл. 3-4).

Таблица 3. Метрики CNN на тестовой выборке (мужские голоса)

	Precision	Recall	F1-score	Accuracy	Кол-во записей
Взрослый	0.77	0.91	0.83	0.76	669
Ребенок	0.72	0.46	0.56		111

Таблица 4. Метрики CNN на тестовой выборке (женские голоса)

	Precision	Recall	F1-score	Accuracy	Кол-во записей
Взрослый	0.76	0.85	0.80	0.73	125
Ребенок	0.64	0.50	0.56		50

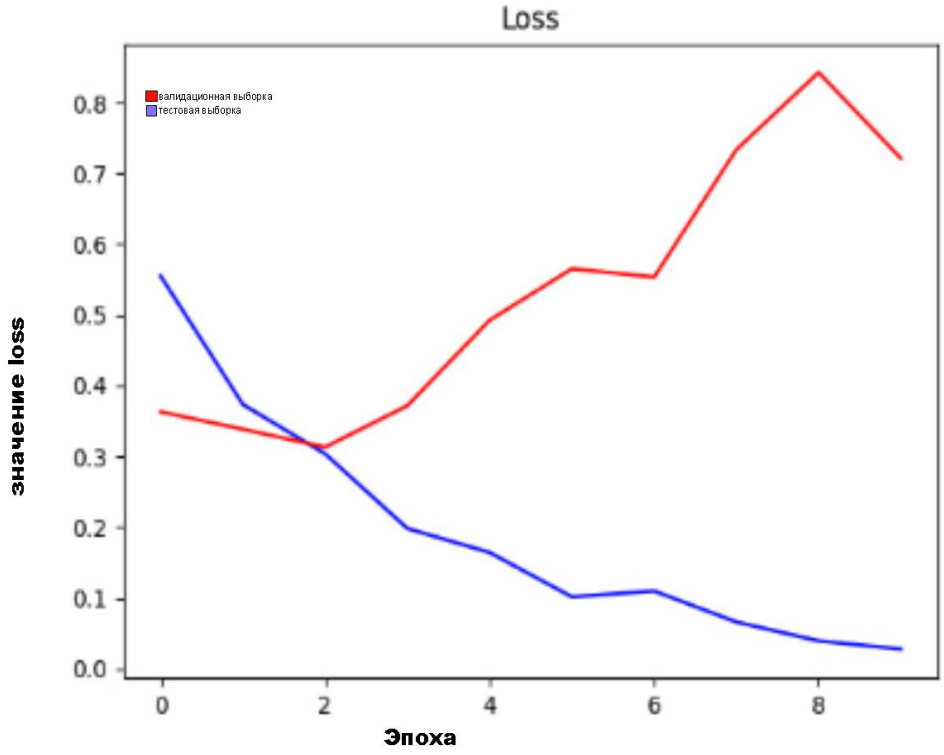

Рис. 4. График функции потерь для CNN

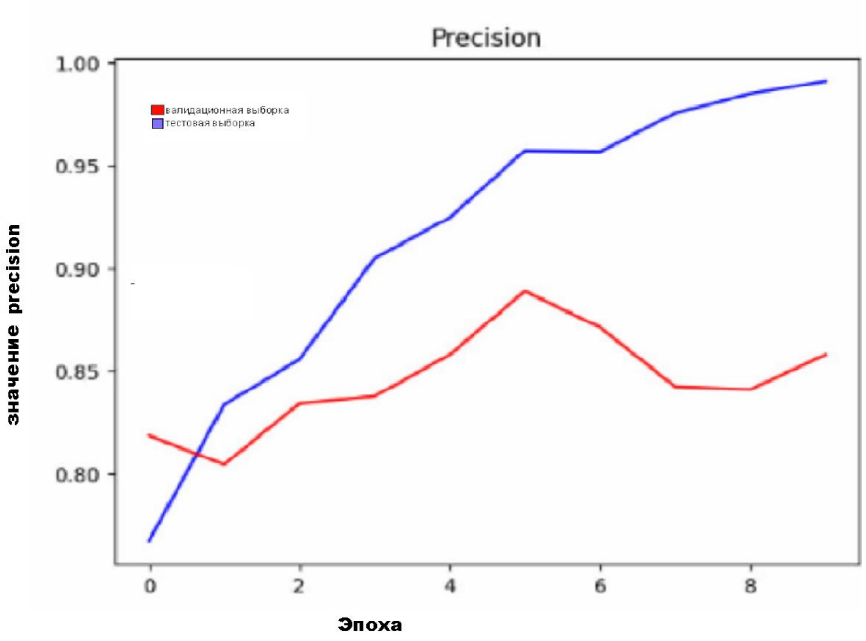

Рис. 5. График precision для CNN

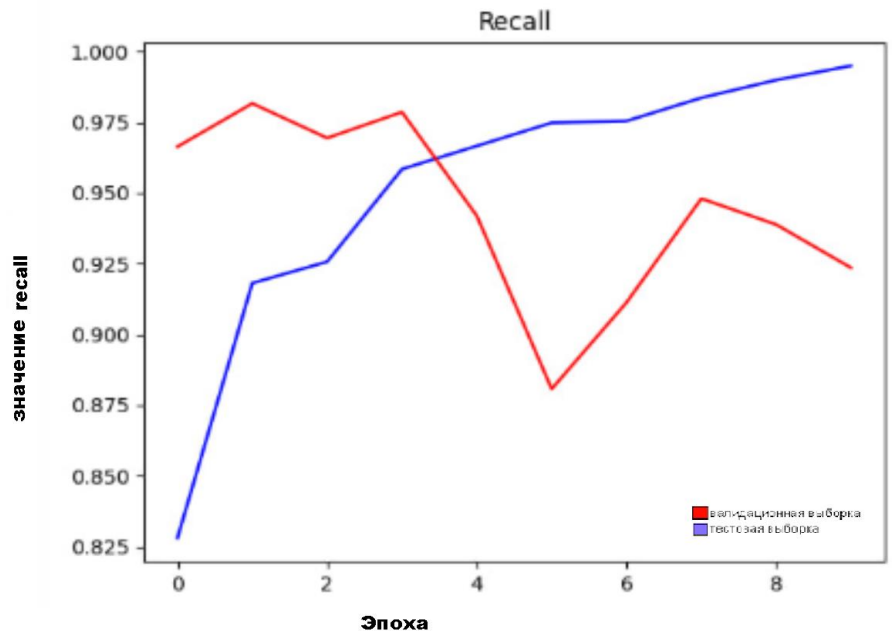

Рис. 6. График recall для CNN

Обучение классификатора на основе эмбеддингов YAMNet

После извлечения 1024-мерных эмбеддингов YAMNet из трёхсекундных аудиофрагментов обучалась модель классификатора. Модель компилировалась с оптимизатором Adam и пониженным темпом обучения 1×10^{-4}, бинарной кросс-энтропией в качестве функции потерь. Обучение велось до 50 эпох с ранней остановкой с параметром patience=10 и мониторингом функции потерь на тестовой выборке. Это позволило автоматически прерывать обучение при отсутствии улучшения и восстанавливать лучшие веса. В результате получалась устойчивая модели с оптимальным балансом между скоростью сходимости и предотвращением переобучения. Результаты представлены в виде метрик accuracy, precision, recall, f1-score и в виде графиков accuracy и функции потерь (рис. 7-8 и табл. 5-6).

Таблица 5. Метрики YAMNet на тестовой выборке (мужские голоса)

	Precision	Recall	F1-score	Accuracy	Кол-во записей
Взрослый	0.95	0.98	0.96	0.94	669
Ребенок	0.84	0.68	0.75		111

Таблица 6. Метрики YAMNet на тестовой выборке (женские голоса)

	Precision	Recall	F1-score	Accuracy	Кол-во записей
Взрослый	0.93	0.89	0.91	0.87	125
Ребенок	0.75	0.82	0.78		50

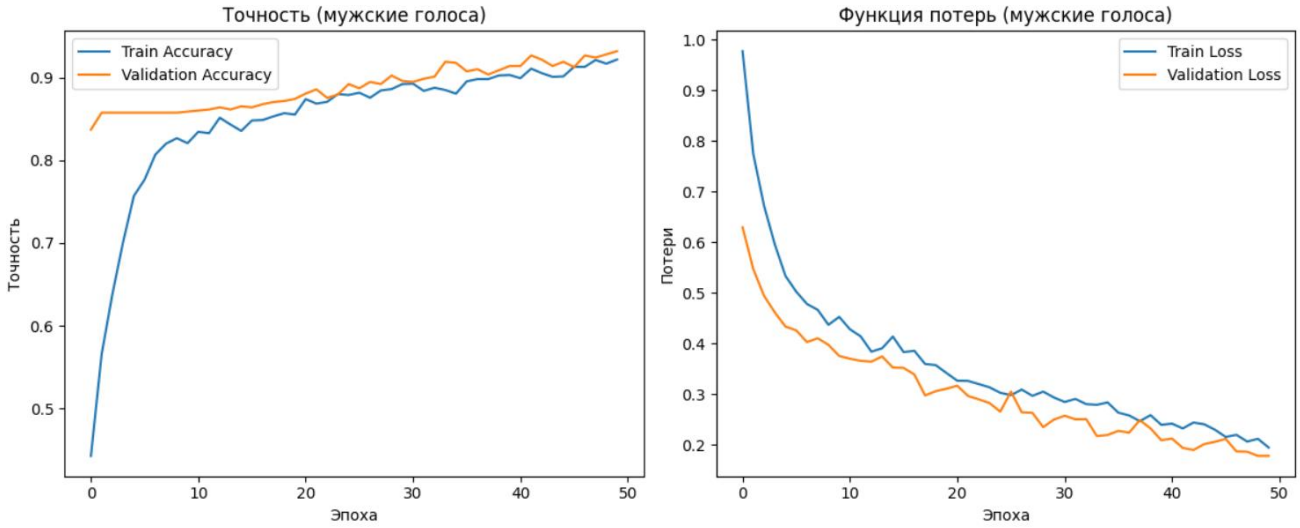

Рис. 7. Accuracy и функция потерь для YAMNet (мужские голоса)

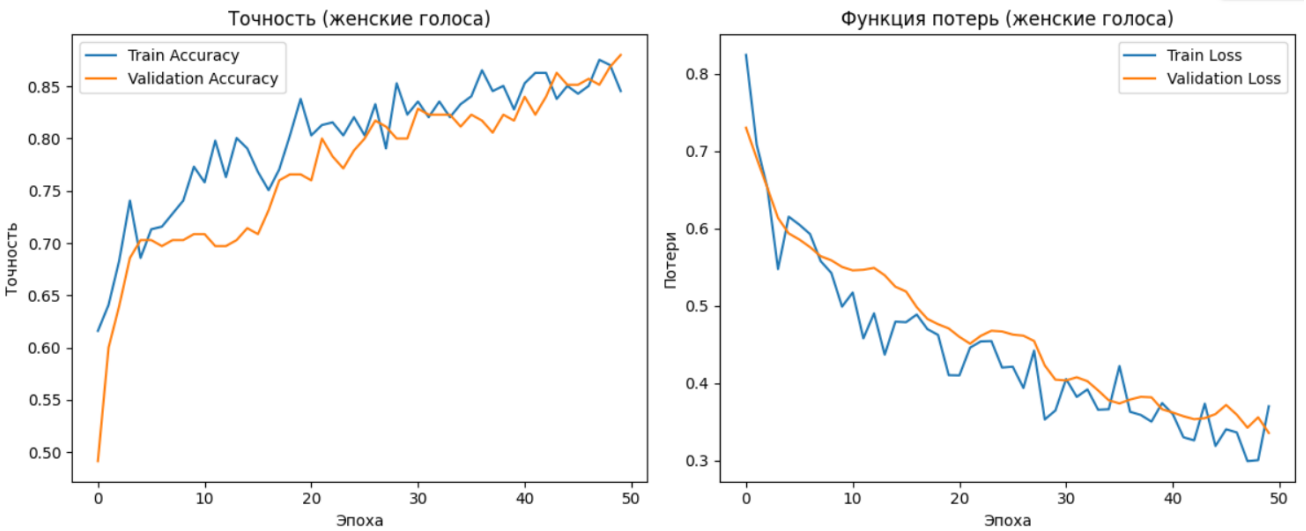

Рис. 8. Accuracy и функция потерь для YAMNet (женские голоса)

Сравнение CNN и YAMNet

Сравнение представлено в сводной таблице метрик (табл. 7).

Таблица 7. Сводная таблица метрик.

Пол	Модель	Класс	Precision	Recall	F1-score	Accuracy
Мужской	CNN	Взрослый	0.77	0.91	0.83	0.76
		Ребенок	0.72	0.46	0.56	
	YAMNet	Взрослый	0.95	0.98	0.96	0.94
		Ребенок	0.84	0.68	0.75	
Женский	CNN	Взрослый	0.76	0.85	0.80	0.73
		Ребенок	0.64	0.50	0.56	
	YAMNet	Взрослый	0.93	0.89	0.91	0.87
		Ребенок	0.75	0.82	0.78	

Заключение

В этом исследовании сравнивались две стратегии определения детского голоса: обучаемая «с нуля» сверточная сеть (CNN) и классификатор на основе предобученных эмбеддингов YAMNet. Эксперименты для мужских и женских голосов показали, что YAMNet устойчиво превосходит CNN по всем ключевым метрикам. Преимущество YAMNet объясняется возможностью извлекать информативные аудио-признаки из трёхсекундных фрагментов благодаря предобучению на большом корпусе AudioSet [4], что согласуется с выводами [6, 10]. Усреднённые эмбеддинги сглаживают шум и сохраняют ключевые характеристики речи, что важно при ограниченных размеченных данных. CNN, в свою очередь, требует большего объёма и однородности тренировочной выборки.

Библиографический список

1. Valcke M. et al. Long-term study of safe Internet use of young children // Computers & Education. – 2011. – T. 57. – P. 1292-1305.

2. Gemmeke J.F. et al. Audio Set: An ontology and human-labeled dataset for audio events // 2017 IEEE ICASSP. IEEE. – 2017. – P. 776-780.

3. Taib D., Tarique M., Islam R. Voice feature analysis for early detection of voice disability in children // 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). – 2018. – P. 12-17.

4. Van Lancker D., Kreiman J., Emmorey K. Familiar voice recognition: Patterns and parameters part I: Recognition of backward voices // Journal of phonetics. – 1985. – T. 13. – P. 19-38.

5. Hershey S. et al. CNN Architectures for Large-Scale Audio Classification // 2017 IEEE ICASSP. IEEE. – 2017. – P. 131-135.

6. Snyder D. et al. X-vectors: Robust DNN embeddings for speaker recognition // 2018 IEEE ICASSP. IEEE. – 2018. – P. 5329-5333.

7. Ardila R. et al. Common Voice: A Massively-Multilingual Speech Corpus // LREC. – 2020. – P. 4218-4222.

8. Wang Y. et al. Tacotron: Towards End-to-End Speech Synthesis // Proceedings of the ISCA InterspeechConference. – 2017.

9. Piczak K.J. Environmental Sound Classification with Convolutional Neural Networks // IEEE International Workshop on MachineLearning for Signal Processing (MLSP). – 2015. – P. 1-6.

10. Baevski A. et al. wav2vec 2.0: A Framework for Self-Supervised Speech Representation Learning // Advances in Neural Information Processing Systems. – 2020. – T. 33. – P. 12449-12460.

APPLICATION OF YAMNET IN CLASSIFICATION OF CHILDREN'S AND ADULT VOICES

A.A. Minsafin, *Graduate Student*
I.M. Kufterin, *Graduate Student*
L.M. Tyshchenko, *Graduate Student*
Supervisor: *A.S. Toshchev, Candidate of Technical Sciences, Associate Professor*
Kazan Federal University
(Russia, Kazan)

Abstract. *The paper presents an experimental comparison of two approaches to the task of automatic child voice detection: a classical convolutional neural network (CNN) and the YAMNet model using pre-trained audio embeddings. For training and testing both methods, a custom dataset was created based on the 'common_voice_17_0' sample (Russian language) considering two key factors – gender (male/female) and age category (children/adults). For YAMNet, data was preprocessed with normalization, truncation or padding to a duration of 3 seconds. In the YAMNet-based approach, the audio signal is transformed into 1024-dimensional vector embeddings, whose temporal mean is used as input for fully connected layers of a multilayer perceptron. In the CNN architecture, the signal is transformed into a mel-coefficient spectrogram and processed by a multilayer convolutional network, then the resulting features are fed to fully connected layers.*

Keywords: *voice recognition; child voice; age classification; convolutional neural networks; CNN; YAMNet.*