УДК 004.032.26
DOI: 10.26102/2310-6018/2025.51.4.004

Исследование задачи автоматизированного сопоставления аудиофайлов

Д.В. Левшин[1]✉, Д.В. Быстряков[1], А.В. Зубков[2]

[1]Волгоградский государственный технический университет, Волгоград, Российская Федерация
[2]Волгоградский государственный медицинский университет, Волгоград, Российская Федерация

Резюме. Объем данных в формате аудиозаписей сильно вырос и продолжает расти, из-за чего с данными становится достаточно сложно работать из-за большого количества различных дубликатов, зашумленных записей, обрезанных записей. В статье представлено решение проблемы поиска нечетких дубликатов аудиозаписей в больших массивах данных. Решение основано на использовании каскадного ансамбля. Для извлечения признаков, анализа временных параметров и оценки сходства между записями использовались сверточные нейронные сети (CNN), сети временных сегментов (TSN), а также сиамские сети. Данные, передаваемые в метод, изначально были преобразованы в изображения формата mel-спектрограмм, созданных с помощью алгоритма кратковременного преобразования Фурье (STFT), то есть каждая аудиозапись нарезалась с определенной частотой дискретизации при условии того, что часть данных имеют связь с предыдущими, преобразовывалась с помощью алгоритма STFT и передавалась в ансамбль моделей. Основное внимание в работе уделено поведению ансамбля с аудиозаписями, которые были подвергнуты различным изменениям, таким как зашумление, искажение, а также обрезка аудиозаписей. Эксперименты, проведенные над набором данных, показали достаточно высокую степень корреляции между результатами, показанными группой людей и методом, что подтверждает эффективность предложенного решения. Метод показал высокую степень устойчивости к различным видам модификации аудиоданных, таких как изменение темпа, добавление шума, а также обрезка аудиозаписей. Дальнейшие исследования могут быть направлены на адаптацию ансамбля к различным типам данных, включая видео и графические записи, что расширит область применения предложенного решения.

Ключевые слова: аудиодубликаты, сверточные сети, преобразование Фурье, аудиошум, устойчивость модели, мел-спектрограмма, сиамская архитектура, временные признаки, сравнение аудиозаписей.

Для цитирования: Левшин Д.В., Быстряков Д.В., Зубков А.В. Исследование задачи автоматизированного сопоставления аудиофайлов. *Моделирование, оптимизация и информационные технологии.* 2025;13(4). URL: https://moitvivt.ru/ru/journal/pdf?id=1903 DOI: 10.26102/2310-6018/2025.51.4.004

Study of the problem of automated matching of audio files

D.V. Levshin[1]✉, D.V. Bystryakov[1], A.V. Zubkov[2]

[1]Volgograd State Technical University, Volgograd, the Russian Federation
[2]Volgograd State Medical University, Volgograd, the Russian Federation

Abstract. The volume of audio recording data has significantly increased and continues to grow, which complicates the processing of such data due to the presence of numerous duplicates, noisy recordings, and truncated audio clips. This article presents a solution to the problem of detecting fuzzy duplicates in large-scale audio datasets. The proposed method is based on the use of a cascaded ensemble. For feature extraction, temporal parameter analysis, and similarity evaluation between recordings,

Convolutional Neural Networks (CNN), Temporal Shift Networks (TSN), and Siamese Networks were utilized. The input data were initially converted into mel-spectrogram images using the Short-Time Fourier Transform (STFT) algorithm. Each audio file was segmented at a specific sampling rate, with attention to temporal continuity, transformed using STFT, and then passed through the ensemble of models. The study focuses on the behavior of the ensemble when processing recordings that have undergone various modifications, such as noise addition, distortion, and trimming. Experiments conducted on the dataset demonstrated a high degree of correlation between the results obtained from human evaluators and the method, confirming the effectiveness of the proposed solution. The method showed strong robustness to different types of audio modifications, such as tempo changes, noise injection, and clipping. Future research may aim to adapt the ensemble to other types of data, including video and graphical recordings, which would expand the applicability of the proposed approach.

Keywords: audio duplicates, convolutional networks, Fourier transform, audio noise, model robustness, mel-spectrogram, siamese architecture, temporal features, comparison of audio recordings.

For citation: Levshin D.V., Bystryakov D.V., Zubkov A.V. Study of the problem of automated matching of audio files. *Modeling, Optimization and Information Technology*. 2025;13(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1903 DOI: 10.26102/2310-6018/2025.51.4.004

Введение

На данный момент человечество сделало значительный шаг вперед в области исследования аудиоданных, и в данной сфере требуется большой объем данных, которые необходимо обработать и подготовить для последующей работы. Их анализ и обработка может занимать часы, недели, а иногда даже месяцы. Определение нечетких дубликатов очень ускорит работу и анализ данных. Нечеткие дубликаты могут возникать из-за изменений качества, темпа, шума или обрезки записей, они приводят к избыточности данных.

Существующие методы, основанные на алгоритмах динамической трансформации или классических сверточных сетях, не всегда справляются с задачей на необходимом уровне и не позволяют достичь высокой скорости обработки данных. В работе предлагается рассмотреть каскадный ансамбль [1], который позволяет объединить преимущества нескольких подходов для повышения точности и производительности за счет использования комбинации моделей.

Для дальнейшего исследования требуется точно определить термин «Нечеткий дубликат». Нечеткий дубликат аудиозаписи – это запись, подвергшаяся искажениям, шумам, обрезке или изменению темпа. Ускоренные и замедленные записи также являются «нечетким дубликатом аудиозаписи». То есть записи не идентичные побитово, а имеющие схожее содержание.

При исследовании аналогичных решений были обнаружены:

– исследование [2], которое демонстрирует практическое применение акустических отпечатков (acoustic fingerprinting) для управления крупными цифровыми архивами аудио, что показывает высокую эффективность при поиске дубликатов в условиях разнообразия источников и качества записей;

– в работе [3] предложен подход с использованием топологического метода персистентная гомология (persistent homology) для извлечения инвариантных признаков звука из мел-спектрограмм, для обеспечения устойчивости к различным видам искажений и шумов, характерных для нечетких дубликатов.

Дополнительные исследования в области представления аудио в виде эмбеддингов, обученных с использованием методов метрического обучения и контрастивных потерь, показывают перспективность глубоких архитектур для решения задач идентификации похожих аудиофрагментов при минимальной зависимости от качества записи. Современные методы ансамблирования моделей показывают более

высокие результаты в точности и скорости определения, что необходимо проверить в результате данного исследования.

Целью данного исследования является разработка и экспериментальная оценка каскадного ансамбля моделей для обнаружения нечетких дубликатов аудиозаписей, для повышения точности и скорости анализа по сравнению с существующими методами. Основное направление работы заключается в использовании моделей нейронной сети, а также адаптации алгоритмов под условия реальных аудиоархивов с разнообразным уровнем шумов, искажений и вариаций в темпе записей.

Задачи в рамках научного исследования можно разбить на следующие:

1) Провести систематический анализ существующих методов обнаружения аудиодубликатов, выявить их преимущества и ограничения, а также определить критерии выбора подходящих методов для решения задачи.

2) Разработать архитектуру каскадного ансамбля моделей и обосновать ее отличие от классических подходов.

3) Определить методику подготовки и преобразования данных, включая использование спектральных и временных признаков для обучения моделей.

4) Провести экспериментальные исследования ансамбля на реальных и синтетически модифицированных аудиоданных, оценив точность, устойчивость и вычислительную эффективность.

5) Сформулировать выводы на основе полученных результатов и определить направления дальнейших исследований.

Формально задачу можно описать следующим образом:

Пусть задано множество аудиозаписей $A = \{a_1, a_2, \dots, a_n\}$. Каждая запись a_i представляется в виде временного сигнала $x_i(t)$: $a_i = x_i(t),\ t \in [0, T_i],\ x_i(t) \in R$. Построим функцию:

$$f: A\ x\ A \to [0,1]. \tag{1}$$

Данная функция для любой пары аудиозаписей (a_i, a_j) возвращает значение степени сходства s_{ij}, удовлетворяющее условиям:

$s_{ij} \approx 1$ если a_i и a_j являются нечеткими дубликатами;

$s_{ij} \approx 0$ если a_i и a_j являются разными записями.

Для этого каждая запись a_i преобразуется в мел-спектрограмму:

$$M_i = MelSpec(a_i) \in R^{FxT}, \tag{2}$$

где F – число мел коэффициентов, T – число временных фреймов. Затем применяется функция признакового отображения:

$$\varphi: R^{FxT} \to R^d. \tag{3}$$

Она реализуется с помощью сверточной нейронной сети и временной сети. Вычисляем итоговое значение:

$$s_{ij} = \sigma\left(-\left|\left|\varphi(M_i) - \varphi\left(M_j\right)\right|\right|_2\right), \sigma \in [0,1], \tag{4}$$

где σ – функция активации. Таким образом, задача поиска нечетких дубликатов аудиозаписей сводится к задаче обучения метрической модели, устойчивой к временным искажениям, шуму и обрезке.

Материалы и методы

На данный момент существует множество решений, основанных на математических методах: сравнение аудиозаписей с помощью деревьев [4], сравнение отпечатков [5], кластерное сравнение [6] и другие, дающие возможность найти нечеткие дубликаты аудиозаписей. К сожалению, сложность этих алгоритмов не позволяет с высокой скоростью обрабатывать большие массивы данных, к тому же существуют более совершенные подходы, где используется машинное обучение.

Рассмотрим самые часто используемые методы для решения задачи более подробно.

Алгоритм динамической трансформации широко применяется в задачах обработки речи и музыки, где временные сдвиги играют ключевую роль. Данный метод позволяет находить оптимальное выравнивание между двумя последовательностями, используя точки совпадения, даже если они отличаются по темпу или длительности. Однако неудовлетворительная вычислительная сложность делает его менее подходящим для работы с большими наборами данных [7, 8].

Если говорить про сверточные сети, то это удобный инструмент определения признаков из фото и видео. Также они отлично выделяют частотные паттерны и звуковые события. В зависимости от способа представления данных, результат может оказаться весьма непредсказуемым. Стоит отметить, что сверточная нейронная сеть имеет ограниченную способность учитывать временные зависимости, поскольку подаются данные в фиксированных окнах. Поэтому модели не обладают достаточной информацией для работы с длинными аудиозаписями с важными временными изменениями [9, 10].

Сиамские сети – эффективный инструмент для сравнения двух объектов или наборов данных, однако требовательные к входным данным. Данные сети обучаются на парах данных, что может быть эффективным методом сравнения пары аудиозаписей. Признаки являются основой обучения текущей модели, и если они не отражают ключевых особенностей аудиозаписи, точность сравнения значительно снижается. Поэтому для сиамских сетей важно использовать мощные методы извлечения признаков [11, 12].

Решение сетей долгосрочной памяти было не один раз представлено для определения результата различных задач поиска и признаков в аудиофайлах [13]. Главная проблема решения заключается в том, что есть много возможных вариантов комбинаций данной сети с другими, так как их самостоятельное решение достаточно сложное и неэффективное.

Спектрограмма – визуальное представление аудиосигнала в виде графика, где по оси X откладывается время, а по оси Y – частота. Интенсивность цвета или яркость показывает амплитуду сигнала на каждой частоте в каждый момент времени. Спектрограмма позволяет анализировать частотные компоненты аудиофайла, что помогает выявить особенности, такие как тембр, тональность и гармоники.

Метод MFCC – это один из самых популярных подходов преобразования аудио, который используется в задачах обработки речи и музыки. Его специфика заключается в извлечении ключевых характеристик звука, которые наиболее важны для восприятия человеком. MFCC учитывает восприятие звука человеком, преобразуя аудио в набор коэффициентов, которые представляют его спектральные особенности.

Метод автокорреляции позволяет измерить степень сходства сигнала с самим собой на различных временных задержках. Зачастую он полезен для поиска повторяющихся паттернов в аудио и может быть использован для выявления схожих фрагментов в аудиофайле, несмотря на изменения в других частях записи.

Векторные представления, или эмбеддинги, позволяют преобразовывать аудиофайлы в многомерные векторы, которые потом могут быть использованы для вычисления сходства между файлами.

В данной работе для преобразования аудиозаписей мы будем использовать метод мел-спектрограмм, так как он:

1) несет малые потери информации при преобразовании и имеет более высокую точность при анализе сильных искажений;

2) позволяет работать с музыкой, звуками и шумами.

Указанные выше причины послужили поводом для выбора данного метода в решении текущей задачи. После выбора алгоритма, каждая запись для сравнения представляется в виде набора мел-спектрограмм. Пример отрезка аудиозаписи представлен на Рисунке 1.

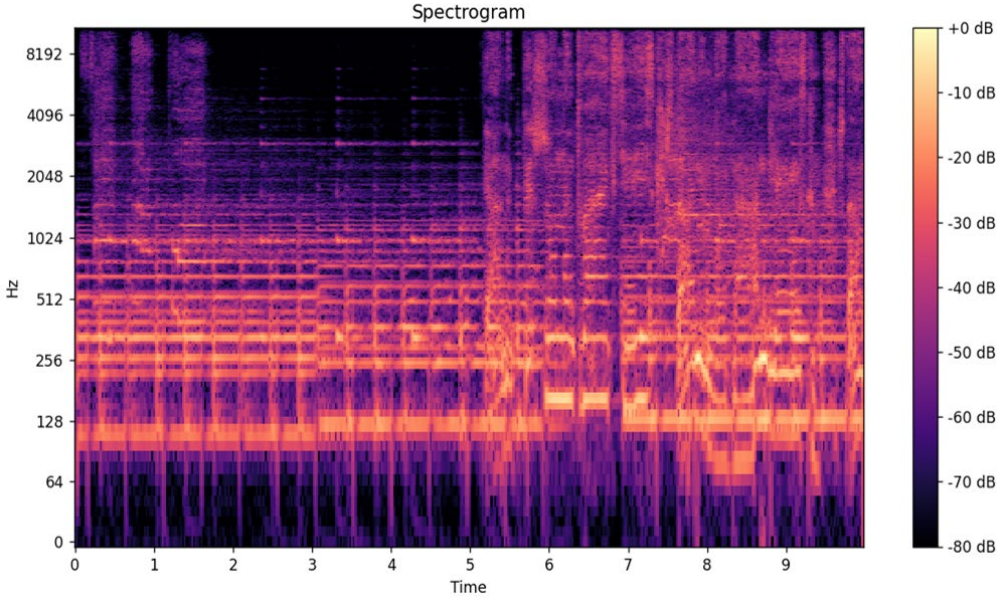

Рисунок 1 – Mel-спектрограмма аудиозаписи
Figure 1 – Mel-spectrogram of the audio recording

После выбора алгоритма предобработки аудиозаписей следует выбор моделей. Предлагаемый метод решения является каскадом трех моделей, первой из них является сверточная нейронная сеть или CNN, которую часто используют из-за эффективного извлечения информативных и устойчивых признаков из входных данных. Она позволяет для каждого изображения создавать вектор признаков [14], которые в дальнейшем передаются в следующую модель каскада.

Следующим компонентом являются сети временных сдвигов (Temporal Shift Networks или TSN). Рассмотрим более подробно принцип работы.

Сначала, каждая аудиозапись преобразовывается в спектрограмму и подается в заранее обученную CNN, где она извлекает ключевые спектральные признаки сигнала. В результате, для каждой аудиозаписи формируется эмбеддинг фиксированной размерности 1×10×512, содержащий локальные частотно-временные характеристики сигнала.

Затем на втором этапе обучения полученные эмбеддинги передаются в модель TSN, предназначенную для агрегации и обобщения временных признаков. TSN преобразовывает каждый тензор признаков CNN в вектор размерности 1×256, содержащий компактное представление аудиозаписи.

После завершения этапов извлечения признаков формируется обучающий датасет для сиамской нейронной сети. Для этого из набора TSN-эмбеддингов конструируются пары: положительные, соответствующие нечетким дубликатам, и отрицательные, представляющие различные по содержанию аудиозаписи. Сиамская сеть обучается на этих парах, чтобы в дальнейшем различать дубликаты и оригиналы, сравнивая расстояния между векторами.

Затем эти модели анализируют последовательности спектрограмм, позволяя выявлять ключевые временные паттерны в аудиозаписи. В результате, аудиофрагмент, представленный как набор спектральных признаков, преобразуется в вектор фиксированной длины, который отражает как временные, так и частотные характеристики сигнала.

Последняя сеть, используемая в данной архитектуре, сиамская, которая служит для сравнения двух аудиозаписей, представленных в форме векторов, с учетом временных характеристик. Эта модель принимает на вход два вектора и обучается различать степени их схожести, используя расстояние между признаками. В результате сиамская сеть формирует окончательную оценку степени сходства между двумя аудиозаписями.

На основании предобработки данных в мел-спектрограмму [15] и использовании каскада моделей была построена архитектура системы, представленная на Рисунке 2.

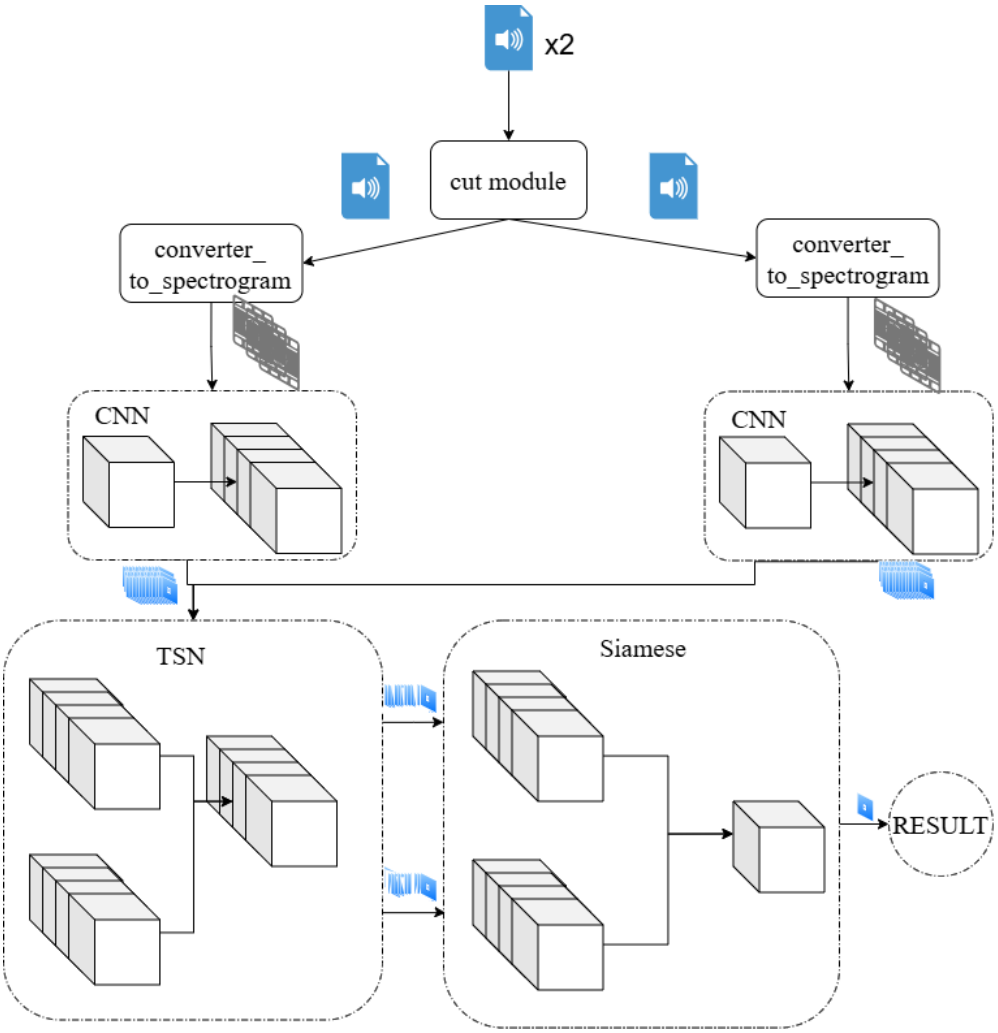

Рисунок 2 – Архитектура системы
Figure 2 – Architecture of the System

На вход подаются два аудиофайла. Модуль нарезки и преобразования аудио в спектрограммы переводит полученные файлы в массив изображений, после чего они подаются в сверточную нейронную сеть, которая преобразует представленные изображения в набор векторов. Следующим этапом сеть временного анализа формирует общий вектор признаков, который передается на сравнение в сиамскую сеть, которая и возвращает результат сравнения аудиофайлов. Данная архитектура позволит достичь высокой точности и скорости поиска нечетких дубликатов аудиозаписей.

Для обучения модели был выбран датасет GTZAN на основе анализа существующих подходов. Он оказался оптимальным выбором для решения поставленной задачи, так как охватывает как популярные музыкальные жанры, так и менее распространенные, такие как «lo-fi» и «ambient». Это дает модели возможность лучше различать аудиофайлы, даже если между ними есть лишь незначительные отличия, как в случае с нечеткими дубликатами.

Выбранный датасет содержит 1000 аудиотреков, каждый длительностью 30 секунд. Он включает по 100 треков для каждого из 10 жанров.

Результаты и обсуждение

После обучения модели и анализа полученных результатов порог для выбора нечетких аудиозаписей был установлен в диапазоне от 70 до 90 %, этот порог позволит избежать излишних выбросов и повысит точность определения необходимых результатов.

Для анализа работы метода был проведен эксперимент: выбрана аудиозапись, и для нее найдены созданные ремиксы. Для этой же аудиозаписи были искусственно созданы следующие нечеткие дубликаты: аудио с искаженным темпом, записи с добавленным шумом и отфильтрованными частотами.

Для повышения объективности эксперимента были привлечены независимые участники. Задачей участников было определить, насколько схожи аудиозаписи. Сперва проигрывалась оригинальная аудиозапись, а после – созданные нечеткие дубликаты. Оценка проводилась по шкале от 1 до 10, где 1 – сходство отсутствует, а 10 – точно является дубликатом.

С помощью мел-спектрограммы можно визуально отобразить сходство записей (Рисунки 3–5). Наиболее сложный случай для распознавания модели – зашумленная аудиозапись (Рисунок 4). Визуализированные результаты представляют собой 10-секундный отрезок записи из набора данных для обучения модели.

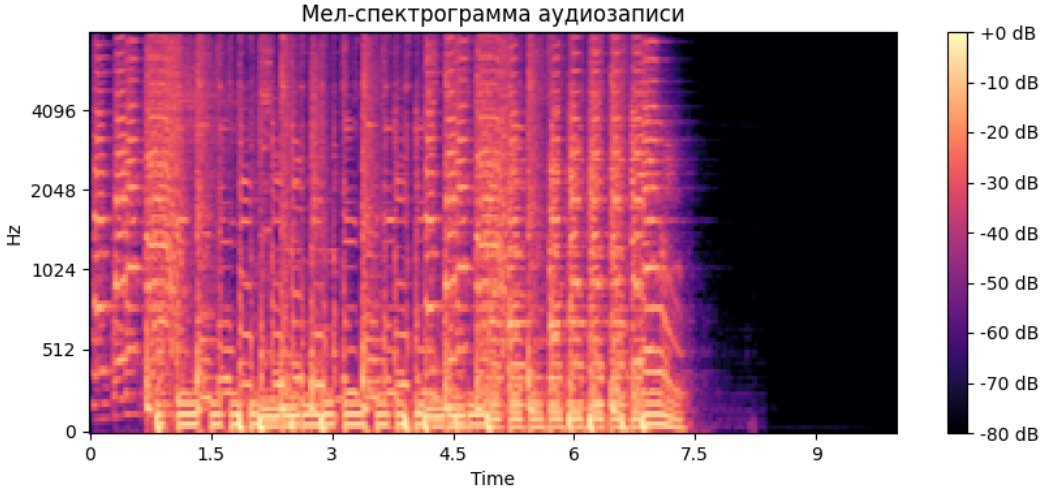

Рисунок 3 – Оригинал аудиозаписи, мел-спектрограмма
Figure 3 – Original audio recording, mel spectrogram

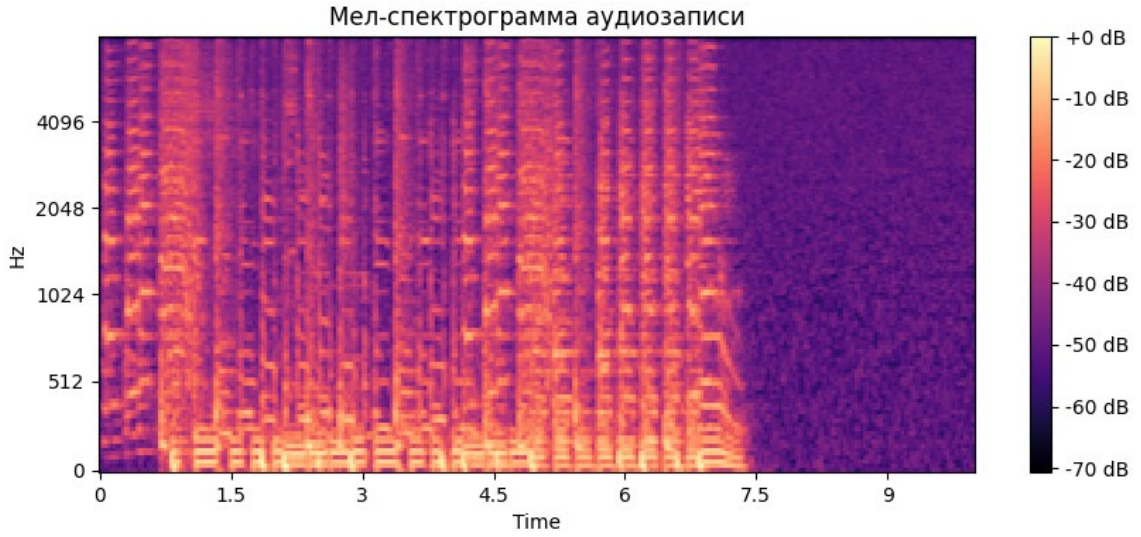

Рисунок 4 – Запись с шумами, мел-спектрограмма
Figure 4 – Recording with noise, mel spectrogram

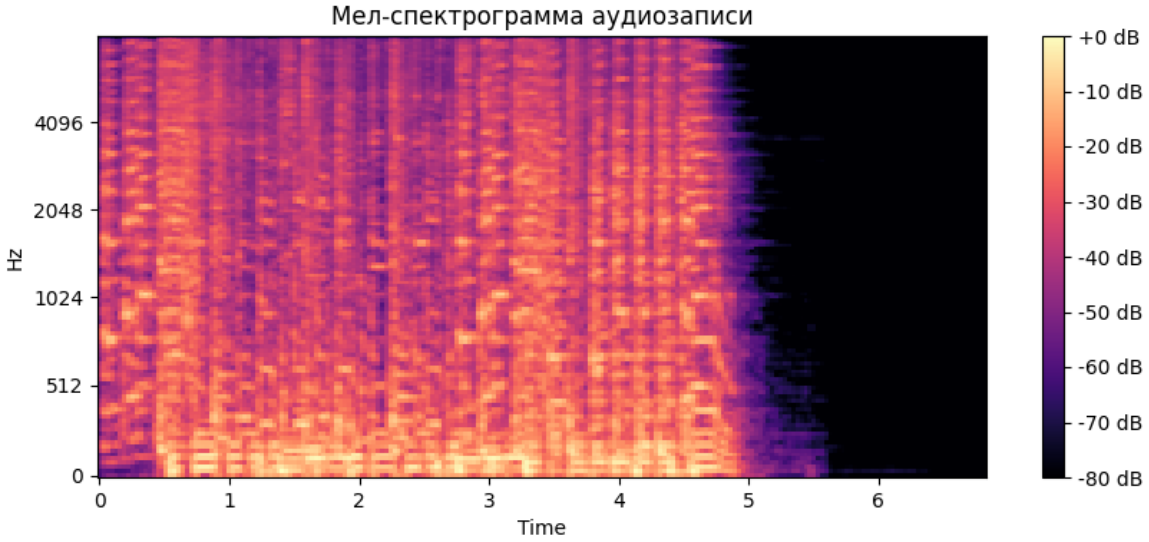

Рисунок 5 – Запись ускоренная, мел-спектрограмма
Figure 5 – Accelerated recording, mel spectrogram

Таблица 2 – Сравнение результатов работы метода и групп людей
Table 2 – Comparison of the results of the method and human groups

	Г1	**Г2**	**Г3**	**Г4**	**Г5**	**М**
Аудиозапись 1 искажение	0,76	0,85	0,7	0,8	0,9	0,8
Аудиозапись 1 ускорение	0,7	0,8	0,8	0,9	0,7	0,75
Аудиозапись 1 шумы	0,5	0,8	0,6	0,7	0,8	0,68
Аудиозапись 2 совершенно	0,01	0,0	0,0	0,0	0,2	0,2
Аудиозапись 2 искажение	0,01	0,0	0,0	0,0	0,0	0,1
Аудиозапись 2 ускорение	0,1	0,0	0,0	0,0	0,1	0,05
Аудиозапись 2 шумы	0,0	0,0	0,0	0,0	0,0	0,15

Опыт проводился следующим образом: было взято 50 человек из разных возрастных групп (7–10 лет, 10–17 лет, 18–25 лет, 26–35 лет, 36+). Каждой группе, включающей от 7 до 10 человек, индивидуально, (Г1–Г5) необходимо было произвести оценку сходства аудиозаписи 1 и остальных искаженных, а также иных форм записей по

10-балльной шкале, после чего эти данные были переведены в систему оценивания от 0 до 1. Сравнительная таблица (Таблица 2) показывает, на сколько процентов ответы метода (М) отличаются от оценки групп (средней по каждому возрастному ограничению).

Для оценки точности метода в сравнении с ответами людей были использованы коэффициенты Спирмена [16] (формула (5)) и Пирсона [17] (формула (6)).

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (5)$$

где r_s – значение коэффициента Спирмена от −1 до 1, отражающее степень монотонной зависимости между переменными; n – количество наблюдений; d_i – разность рангов между переменными для каждой пары.

$$r_p = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\cdot\sqrt{\sum(y_i-\bar{y})^2}}, \quad (6)$$

где r_p – коэффициент линейной корреляции Пирсона от −1 до 1. x_i, y_j – значение признаков *x* и y в *i*-й паре. $\bar{x}$ среднее значение по выборке *x* и $\bar{y}$ – среднее значение по выборке *y*.

По результатам оценки группами людей было взято среднее значение и по формуле (5) рассчитан результат коэффициента Спирмена, равный 0,84, что свидетельствует о высоком уровне корреляции между оценками участников и метода. Коэффициент Пирсона составил 0,98, что указывает на наличие линейной зависимости между ответами опрашиваемых участников и метода, подтверждая результаты.

Также для оценки результатов был проведен анализ методов, решающих с помощью других архитектур. Анализ сравнения приведен в Таблице 3. Для сравнения используется набор из 1000 песен в 4 различных типах нечетких дубликатов (Таблица 3).

Таблица 3 – Сравнение результатов работы аналогов и модели
Table 3 – Comparison of the results of analogs and models

Метод	Точность, %	Скорость на 1000 записей, с	Проблемы
1	2	3	4
Сиамские сети, STFT	Нарезка: 54 Эхо: 63 Реверс: 95 Шумы: 33	630	Умеренная точность при нарезке и шуме.
Сиамские сети, MFCC	Нарезка: 60 Эхо: 68 Реверс: 93 Шумы: 40	680	Улучшенная устойчивость к шумам по сравнению с STFT.
CNN, STFT	Нарезка: 54 Эхо: 63 Реверс: 95 Шумы: 50	500	Неустойчивость к шумам и нарезке. Слабо справляется с эхо.
CNN, MFCC	Нарезка: 34 Эхо: 83 Реверс: 95 Шумы: 20	580	Неустойчивость к шумам и нарезке.

Таблица 3 (продолжение)
Table 3 (continued)

LSTM, MFCC	Нарезка: 74 Эхо: 68 Реверс: 95 Шумы: 56	900	Длительное время обработки. Точность эхо.
LSTM, STFT	Нарезка: 74 Эхо: 63 Реверс: 82 Шумы: 55	630	Умеренная точность при реверсе.
STFT, DTW	Нарезка: 6 Эхо: 33 Реверс: 95 Шумы: 33	1200	Низкая точность при нарезке и эхо. Высокое время обработки.
Акустический фингерпринтинг	Нарезка: 50 Эхо: 63 Реверс: 10 Шумы: 20	100	Чувствителен к шумам и реверсу.
Предложенный метод	Нарезка: 85 Эхо: 96 Реверс: 78 Шумы: 42	350	Высокая точность при нарезке и эхо. Умеренная точность при реверсе. Быстрое время обработки.

Предлагаемый метод представляет собой оптимальное решение для задачи обнаружения нечетких дубликатов аудиозаписей. Он сочетает в себе высокую точность при различных искажениях и быстрое время обработки, превосходя другие методы по данным показателям.

Заключение

Таким образом, поставленная задача была формализована, исследованы самые распространенные для решения установленной задачи методы и выявлены наиболее подходящие, которые позволяют повысить скорость и точность определения нечетких дубликатов аудиозаписей. Выбранный способ заключается в использовании каскадного подхода из сверточной нейронной сети, сети для анализа временных рядов и сиамской сети для сравнения векторов признаков. Это решение сочетает в себе высокую точность и скорость за счет использования легковесных моделей на ранних этапах фильтрации, что делает его особенно полезным при работе с большими массивами аудиоданных.

Проведение тестирования на группе лиц разных возрастных категорий показало высокий уровень корреляции между результатами модели и данных групп. Коэффициент Спирмена составил 0,84, а коэффициент Пирсона 0,98, что свидетельствует о наличии значимой линейной зависимости. Также сравнение с аналогами показало, что модель справляется с нечеткими дубликатами типа «нарезка» с точностью в 85 %, с задачей определения эхо с точностью 96 %, реверсные записи с точностью в 78 %. Слабым участком определения являются шумы, однако точность их определения среди аналогов имеет также значение выше среднего и составляет 42 %.

Предложенный метод демонстрирует потенциал не только в области распознавания нечетких дубликатов аудиозаписей, но и в более широком спектре задач мультимедийного анализа, включая обработку видео и изображений. Такое расширение открывает возможности для создания универсальных программных решений в различных прикладных областях.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Елена Алексеевна Кочегурова, Софья Михайловна Сайберт, Ксения Витальевна Татьянкина Оптимизация параметров гибридного алгоритма прогнозирования с использованием ансамбля моделей в режиме реального времени // Известия Томского политехнического университета. Промышленная кибернетика. 2024. №4. URL: https://cyberleninka.ru/article/n/optimizatsiya-parametrov-gibridnogo-algoritma-prognozirovaniya-s-ispolzovaniem-ansamblya-modeley-v-rezhime-realnogo-vremeni (дата обращения: 11.04.2025).
2. Six, J., Bressan, F., Renders, K. (2023). Duplicate Detection for for Digital Audio Archive Management: Two Case Studies. In: Biswas, A., Wennekes, E., Wieczorkowska, A., Laskar, R.H. (eds) Advances in Speech and Music Technology. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-18444-4_16
3. Reise W., Fernández X., Dominguez M., Harrington H. A., Beguerisse-Díaz M. Topological fingerprints for audio identification [Electronic resource] // arXiv preprint. 2023. No. 2309.03516. URL: https://arxiv.org/abs/2309.03516 (дата обращения: 31.07.2025).
4. Маленко, С. А. Увеличение производительности алгоритмов поиска дубликатов аудиозаписей / С. А. Маленко. — Текст: непосредственный // Молодой ученый. — 2017. — № 49 (183). — С. 22-26. — URL: https://moluch.ru/archive/183/47026/ (дата обращения: 11.04.2025).
5. Ryynanen, Matti & Klapuri, Anssi. (2008). Query by humming of MIDI and audio using locality sensitive hashing. Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2249 - 2252. 10.1109/ICASSP.2008.4518093.
6. Булавин Д.А., Харитонов И.А. Анализ методов распознавания и преобразования аудиоинформации в ноты // Автоматизированные системы управления и приборы автоматики. — 2011. — № 152(2). — С. 56–63. — URL: https://goo.su/023eT4W (дата обращения: 11.04.2025).
7. Новохрестова Д.И. Временная нормализация слогов алгоритмом динамической трансформации временной шкалы при оценке качества произнесения слогов// Компьютерные и информационные науки// URL: https://goo.su/FWWMo8E (дата обращения: 11.04.2025).
8. Wang, Y., Lyu, X. & Yang, S. Ocean observing time-series anomaly detection based on DTW-TRSAX method. J Supercomput 80, 18679–18704 (2024). https://doi.org/10.1007/s11227-024-06183-w
9. Ustubioglu, A., Ustubioglu, B. & Ulutas, G. Mel spectrogram-based audio forgery detection using CNN. SIViP 17, 2211–2219 (2023). https://doi.org/10.1007/s11760-022-02436-4
10. Zhao, H., Ye, Y., Shen, X. et al. 1D-CNN-based audio tampering detection using ENF signals. Sci Rep 14, 11186 (2024). https://doi.org/10.1038/s41598-024-60813-0
11. Wang, W., Lu, Z. Few-shot bronze vessel classification via siamese fourier networks. Sci Rep 14, 18011 (2024). https://doi.org/10.1038/s41598-024-69272-z
12. Lin, YB., Bertasius, G. (2025). Siamese Vision Transformers are Scalable Audio-Visual Learners. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds) Computer Vision – ECCV 2024. ECCV 2024. Lecture Notes in Computer Science, vol 15072. Springer, Cham. https://doi.org/10.1007/978-3-031-72630-9_18

13. India, Miquel & Fonollosa, José & Hernando, Javier. (2017). LSTM Neural Network-Based Speaker Segmentation Using Acoustic and Language Modelling. 2834-2838. 10.21437/Interspeech.2017-407.
14. Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K.. // CNN Architectures for Large-Scale Audio Classification // arXiv. – 2017. – URL: https://arxiv.org/abs/1609.09430 (дата обращения: 11.04.2025).
15. А.С. Ананьев, Д.В. Бутенко, К.В. Попов, Моделирование процессов управления качеством продукции на основе имитационного моделирования // Инженерный вестник Дона. — 2012. — № 2. — URL: http://www.ivdon.ru/ru/magazine/archive/n2y2012/815 (дата обращения: 11.04.2025).
16. Кошелева Н. Н. Корреляционный анализ и его применение для подсчета ранговой корреляции Спирмена // Актуальные проблемы гуманитарных и естественных наук. 2012. №5. URL: https://goo.su/0zy6O (дата обращения: 11.04.2025).
17. Меньшов М. Коэффициент корреляции Пирсона. – Кафедра математической статистики ИВМиИТ КФУ 2020г. – Режим доступа: https://kpfu.ru/portal/docs/F_2064674290/NPS_19.Pirson.Menshov.pdf (дата обращения: 11.04.2025).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Левшин Денис Витальевич, магистрант, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: levshin01@bk.ru
ORCID: 0009-0001-5163-1393

Denis V. Levshin, Master's Degree student, Volgograd State Technical University, Volgograd, the Russian Federation.

Быстряков Даниил Владимирович, магистрант, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: bystriackoff@yandex.ru
ORCID: 0009-0004-0391-3849

Daniil V. Bystryakov, Master's Degree student, Volgograd State Technical University, Volgograd, the Russian Federation.

Зубков Александр Владимирович, кандидат технических наук, доцент кафедры программного обеспечения автоматизированных систем Волгоградского государственного технического университета, Волгоград, Российская Федерация.
e-mail: aleksandr.zubkov@volgmed.ru
ORCID: 0000-0003-0425-5695

Alexander V. Zubkov, Candidate of Engineering Sciences, Associate Professor at the Department of Software for Automated Systems, Volgograd State Technical University, Volgograd, the Russian Federation.

Статья поступила в редакцию 17.04.2025; одобрена после рецензирования 09.09.2025; принята к публикации 25.09.2025.

The article was submitted 17.04.2025; approved after reviewing 09.09.2025; accepted for publication 25.09.2025.