ISSN 2782-2001
Системы анализа и обработки данных
том 99, № 3, 2025, с. 69–82

http://journals.nstu.ru/vestnik
Analysis and data processing systems
Vol. 99, No. 3, 2025, pp. 69–82

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ

INFORMATION TECHNOLOGIES AND TELECOMMUNICATIONS

УДК 004.032.26 DOI: 10.17212/2782-2001-2025-3-69-82

Применение сверточных и сиамских нейронных сетей в задаче сравнения нечетких дубликатов аудиообъектов*

Д.В. ЛЕВШИН[a], Д.В. БЫСТРЯКОВ[b], М.А. СКЛЯРОВ[c], А.В. ЗУБКОВ[d]

400005, Россия, г. Волгоград, проспект им. В.И. Ленина, 28, Волгоградский государственный технический университет

[a] *levshin01@bk.ru* [b] *bystriackoff@yandex.ru* [c] *mikhail.29.06.2001@gmail.com*
[d] *aleksandr.zubkov@volgmed.ru*

В настоящее время наблюдается рост объема данных в формате аудиозаписей, с которыми достаточно сложно работать из-за большого количества дубликатов, зашумленных или обрезанных записей. В статье рассматривается один из вариантов решения проблемы поиска нечетких дубликатов аудиозаписей в больших массивах данных. Предлагаемое решение основано на использовании каскадного ансамбля моделей для определения нечетких дубликатов. Для извлечения признаков, анализа временных параметров и оценки сходства между записями использовались сверточные нейронные сети (CNN, Convolutional Neural Networks), сети временных сдвигов (TSN, Temporal Shift Networks), а также сиамские нейронные сети. Аудиоданные, подаваемые в модели, предварительно преобразовывались в ряд спектрограмм, созданных с помощью алгоритма кратковременного преобразования Фурье (STFT, Short-Time Fourier Transform). Каждая аудиозапись нарезáлась с заданной частотой дискретизации, преобразовывалась с использованием STFT и передавалась в ансамбль моделей. Основное внимание в работе уделено исследованию поведения ансамбля при работе с аудиозаписями, подвергнутыми различным изменениям, таким как зашумление, искажение или обрезка. Эксперименты, проведенные на наборе данных, продемонстрировали высокую степень корреляции между результатами, полученными группой людей, и результатами, выданными ансамблем моделей, что подтверждает эффективность предложенного подхода. Ансамбль моделей показал высокую устойчивость к различным видам модификаций аудиоданных, таким как изменение темпа, добавление шума и обрезка записей. Дальнейшие исследования могут быть направлены на адаптацию ансамбля к различным типам данных, включая видео- и графические записи, что позволит расширить область применения предложенного метода.

Ключевые слова: нечеткий дубликат, свертка, нейронная сеть, STFT, шумы, производительность, спектрограмма, сиамская сеть, сегментация, каскадный ансамбль

* *Статья получена 17 февраля 2025 г.*

ВВЕДЕНИЕ

Для дальнейшего исследования необходимо точно определить термин «нечеткий дубликат аудиозаписи». Под ним понимаются аудиофайлы, представляющие собой модифицированные версии одной и той же записи: обрезанные, зашумленные, ускоренные или замедленные. Эти изменения затрудняют автоматическое сопоставление записей и требуют специальных алгоритмов для выявления семантически схожих, но технически отличающихся экземпляров.

Актуальность этой задачи обусловлена быстрым ростом объемов аудиоданных, хранимых и анализируемых в различных отраслях – от цифровых архивов и медиабиблиотек до платформ дистанционного обучения и систем распознавания речи. Наличие таких дубликатов приводит к избыточному расходу ресурсов хранения, снижению эффективности обучения моделей и осложнению последующего анализа.

Поиск нечетких дубликатов необходим в ряде практических задач:

1) при очистке датасета перед обучением систем распознавания речи или классификации звуков;

2) в системах контроля авторских прав, где запись может быть загружена с искажением с целью обхода фильтров.

Несмотря на наличие современных методов анализа, включая алгоритмы динамического временного выравнивания (DTW) [1] и сверточные нейронные сети (CNN), не всегда обеспечивается требуемая точность и производительность при работе с видоизмененными аудиофрагментами. В связи с этим в настоящей работе предлагается использовать каскадный ансамбль моделей, объединяющий несколько подходов, что позволяет повысить точность и устойчивость к разнообразным искажениям.

1. ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ ДЛЯ ПОИСКА НЕЧЕТКИХ ДУБЛИКАТОВ

Существует ряд подходов к поиску нечетких дубликатов аудиозаписей, основанных на математических методах, таких как сравнение аудиозаписей с помощью деревьев [2], сравнение аудиоотпечатков [3], кластерный анализ [4] и другие. Однако высокая вычислительная сложность этих алгоритмов затрудняет их применение для обработки больших массивов данных. Кроме того, существуют более эффективные по скорости и точности решения, основанные на методах машинного обучения и нейронных сетях, среди которых можно выделить следующие.

1. Алгоритм динамического преобразования (Dynamic Time Warping, DTW) – метод, широко применяемый в задачах обработки речи и музыки, где временные искажения играют ключевую роль. Он позволяет находить оптимальное выравнивание между двумя последовательностями с учетом точек совпадения, даже если они различаются по темпу или длине. Однако высокая вычислительная сложность DTW ограничивает его применение при работе с большими наборами данных [5, 6].

2. Сверточные нейронные сети (CNN) – мощный инструмент для извлечения признаков из изображений и видео. В контексте аудиоряда CNN эффек-

тивно выявляют частотные паттерны и звуковые события, особенно при соответствующем представлении аудиоданных, например, в виде спектрограмм. Однако такие сети не лучшим образом учитывают временные зависимости, так как анализируют данные в фиксированных временных окнах. Это снижает их эффективность при работе с длинными аудиозаписями, где временной контекст играет важную роль [7, 8].

3. Сиамские нейронные сети – эффективный подход для сравнения двух объектов или наборов данных. Эти модели обучаются на парах примеров и могут точно определять степень сходства между аудиозаписями. Однако эффективность сиамских сетей напрямую зависит от качества извлеченных признаков. Если признаки не отражают ключевых характеристик аудиозаписей, точность сравнения существенно снижается. Поэтому для таких моделей важно использовать надежные методы предварительной обработки и извлечения признаков [9, 10].

2. ВЫБОР АЛГОРИТМА ПРЕДОБРАБОТКИ АУДИОЗАПИСИ

Использование различных форм представления аудиозаписей позволяет повысить точность работы моделей, поэтому выбор метода предобработки данных является одним из ключевых этапов.

Для повышения точности в настоящей работе выбран алгоритм извлечения mel-частотных кепстральных коэффициентов (MFCC) и mel-спектрограммы (табл. 1).

Таблица 1

Table 1

Сравнение эффективности алгоритмов предобработки аудиоданных для решения задачи поиска нечетких дубликатов

Comparison of the efficiency of audio data preprocessing algorithms for solving the problem of finding fuzzy duplicates

Метод	Лучший для	Преимущества	Ограничения
Спектрограммы	Глубокое обучение (CNN, TSN)	Отлично работает с CNN, сохраняет частотную информацию	Потеря точного времени
MFCC	Классические алгоритмы (DTW, SVM)	Компактное представление [11]	Потеря спектральных деталей
Временны́е признаки	Глубокие нейросети (WaveNet, Wav2Vec)	Полное сохранение информации	Высокие требования к данным
Аудиоотпечатки	Быстрый поиск дубликатов	Высокая скорость, низкие ресурсы	Чувствительность к обрезке

После выбора алгоритма каждая запись для сравнения представляется в виде набора mel-спектрограмм, представленных на рис. 1.

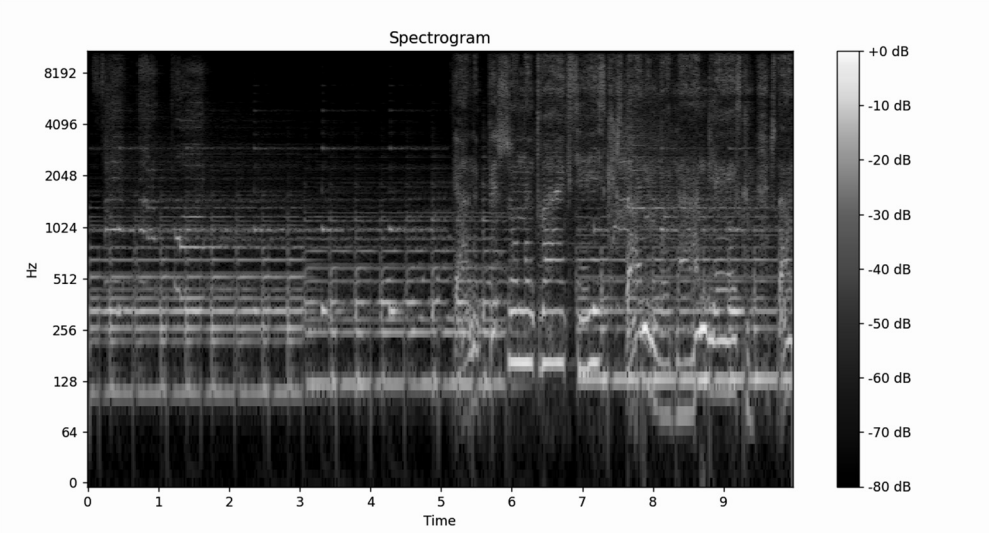

Рис. 1. Аудиозапись, представленная в виде mel-спектрограммы

Fig. 1. An audio recording presented as a mel spectrogram

Теперь стоит вопрос в выборе моделей, используемых для определения нечетких аудиозаписей.

3. АРХИТЕКТУРА

После выбора метода предобработки аудиозаписей следующим этапом является выбор архитектуры моделей. Первым компонентом является сверточная нейронная сеть (CNN), которая используется для преобразования каждого изображения (например, спектрограммы) в вектор признаков [12]. Эти векторы далее передаются в следующую модель.

Следующим этапом обработки служат временные сверточные сети (Temporal Shift Networks), которые анализируют последовательности изображений, извлеченных из аудиозаписей, и выявляют ключевые временные паттерны. В результате одна аудиозапись, представленная в виде последовательности спектрограмм, преобразуется в вектор фиксированной длины, отражающий как спектральные, так и временны́е характеристики сигнала.

Заключительным элементом архитектуры является сиамская нейронная сеть, предназначенная для сравнения двух аудиозаписей, представленных в виде векторов. Эта модель учитывает временны́е характеристики и позволяет получить итоговую оценку степени сходства между двумя аудиозаписями.

На основе выбранных методов предобработки [13] и архитектуры моделей была построена структура системы, представленная на рис. 2.

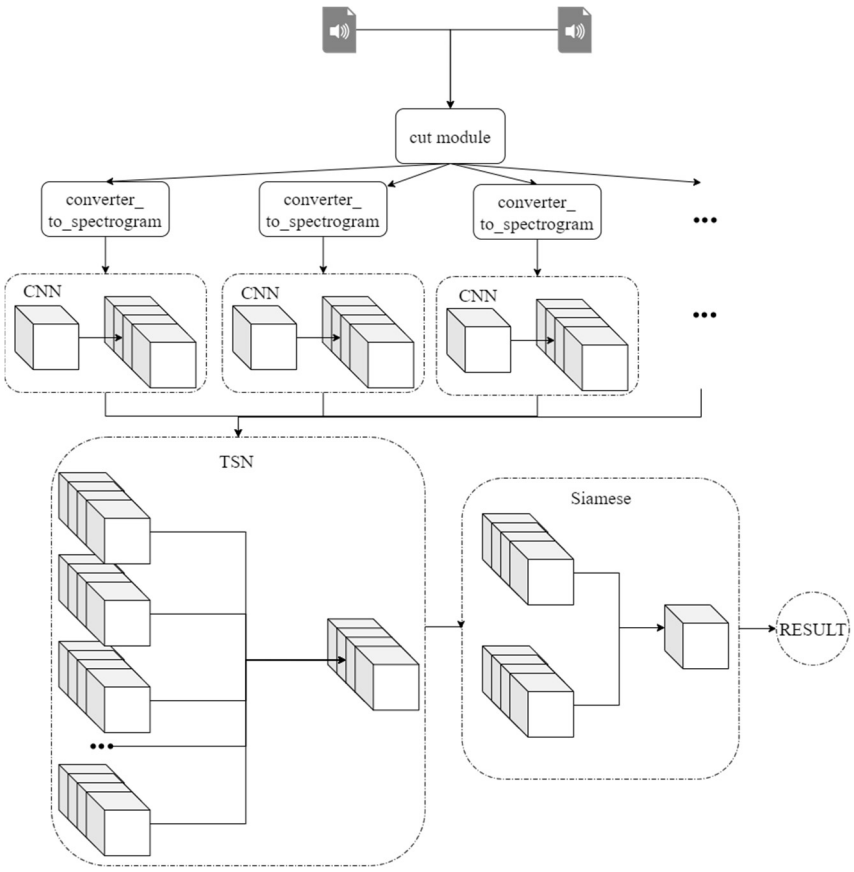

Рис. 2. Архитектура метода

Fig. 2. Architecture of the method

Предложенная архитектура позволяет обеспечить высокую точность и скорость при поиске нечетких дубликатов аудиозаписей. Следующим этапом является сбор данных и обучение ансамбля моделей.

4. ПОДГОТОВКА ДАННЫХ И ОБУЧЕНИЕ

Обучение системы моделей включает следующие этапы.

1. Сбор данных: необходимо сформировать датасет аудиозаписей, используя либо собственные данные, либо открытые источники с уже размеченными аудиофрагментами.

2. Организация данных: данные следует структурировать по папкам, например, выделив отдельные каталоги для дубликатов и уникальных аудиозаписей.

3. Предобработка: загруженные аудиофайлы преобразуются в mel-спектрограммы, которые являются входными данными для последующих моделей [14].

4. Обучение модели CNN: с помощью сверточной нейронной сети из каждой спектрограммы извлекаются векторы признаков, которые затем используются для обучения временной модели.

5. Обучение модели TSN: временная сверточная сеть обучается на последовательностях векторов, извлеченных CNN, формируя итоговое представление каждой аудиозаписи.

6. Обучение сиамской сети: на последнем этапе производится обучение сиамской нейросети на парах векторов в формате [вектор 1, вектор 2, метка], где метка указывает, являются ли аудиозаписи дубликатами.

5. АНАЛИЗ РЕЗУЛЬТАТОВ, СТЕПЕНЬ СХОДСТВА

В рамках исследования был осуществлен анализ результатов работы обученного каскада моделей, который позволил выявить следующие закономерности в полученных данных.

1. Степень сходства между двумя полностью идентичными аудиозаписями 92…100 %.

2. Для совершенно различных записей показатель сходства составляет 0,01…35 %.

3. В редких случаях наблюдается сходство порядка 55…58 % между двумя аудиозаписями.

Для проведения оценки качества каскада необходимо определить порог сходства двух аудиозаписей, который будем считать нечеткими дубликатами. На основании данных, полученных при проведении опытов, взят диапазон 60…90 %. Такой выбор позволяет минимизировать количество ложноположительных и ложноотрицательных результатов, тем самым повышая точность системы.

6. ВАЛИДАЦИЯ РАБОТЫ КАСКАДНОГО АНСАМБЛЯ МОДЕЛЕЙ

После обучения модели была поставлена задача оценить ее качество путем проведения социального эксперимента. Для этого был проведен эксперимент, целью которого стало сравнение результатов работы модели с оценками, полученными от группы людей.

В рамках эксперимента была выбрана оригинальная аудиозапись и собраны ее ремиксы. Дополнительно для нее были искусственно созданы нечеткие дубликаты:

1) версии с ускоренным темпом;

2) записи с добавленным шумом;

3) записи с фильтрацией определенных частот.

Все варианты были проанализированы как с помощью модели, так и вручную – участниками эксперимента.

Для повышения объективности в исследовании принимали участие люди, не знакомые с целью эксперимента. Их задачей было оценить степень сходства между оригинальной и искаженной аудиозаписью. Оригинал проигрывался первым, затем – модифицированный вариант. Оценка проводилась по шкале от 1 до 10, где 1 – это полностью отличные, а 10 – полностью идентичные. Эксперимент был проведен на трех различных оригинальных аудиозаписях, и оценки были получены как от участников, так и от модели (ансамбля).

Для начала сравним визуальное представление аудиозаписи 1 и двух ее модифицированных версий – с ускорением и добавлением шума. Для этого были построены изображения на частотных преобразованиях, что показано на рис. 3 и на mel-спектрограммах, которые представлены на рис. 4, 6 и 8.

Каждое изображение основано на 10-секундном фрагменте аудиозаписи. На частотных спектрах различия трудно различимы визуально, в то время как mel-спектрограммы позволяют четко увидеть изменения – ускорение, появление шумов и искажения. Это подтверждает, что mel-спектрограмма является более информативной формой представления аудиосигнала при визуальном анализе.

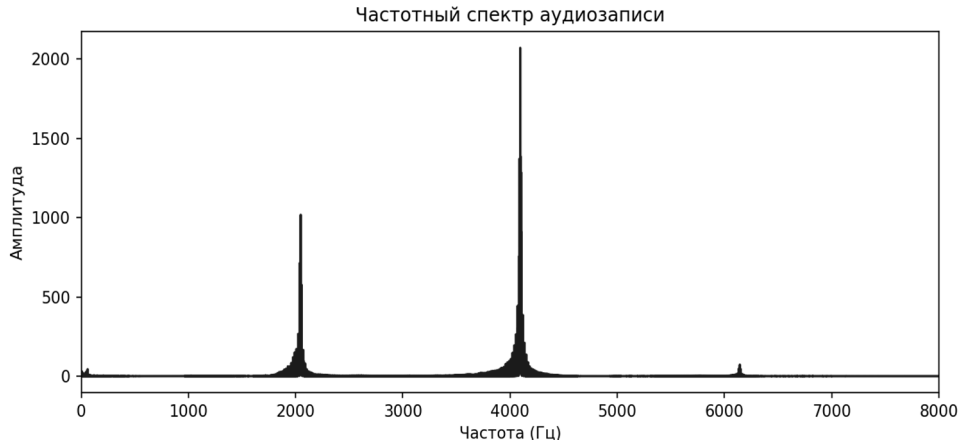

Рис. 3. Оригинал аудиозаписи, частотный спектр

Fig. 3. An original audio recording, a frequency spectrum

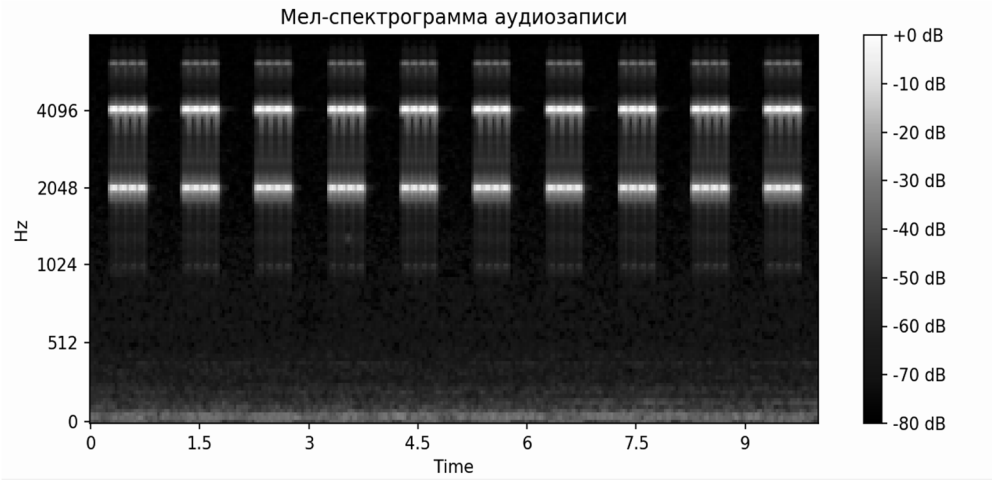

Рис. 4. Оригинал аудиозаписи, mel-спектрограмма

Fig. 4. An original audio recording, a mel spectrogram

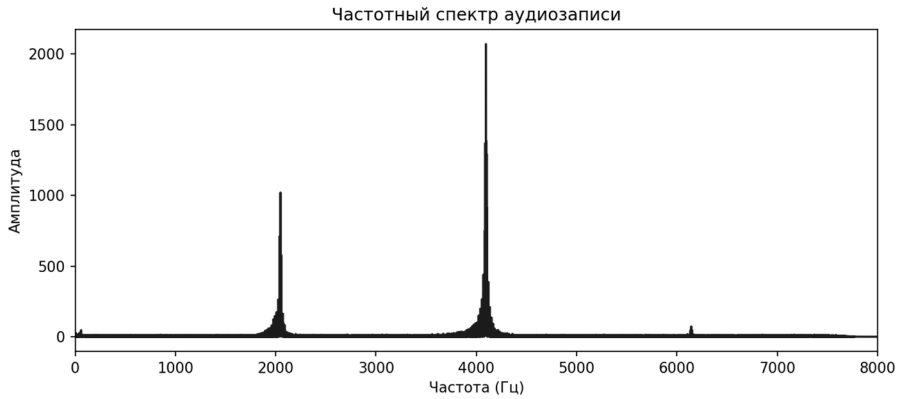

Рис. 5. Запись с шумами, частотный спектр

Fig. 5. A recording with noise, a frequency spectrum

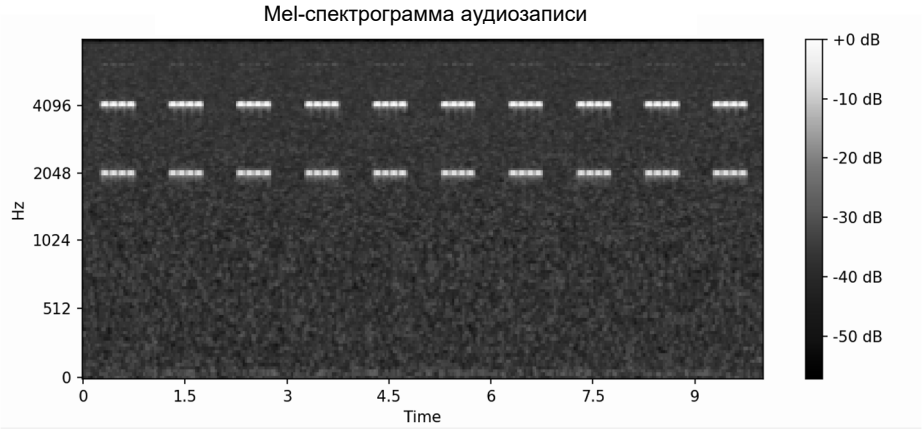

Рис. 6. Запись с шумами, mel-спектрограмма

Fig. 6. A recording with noise, a mel spectrogram

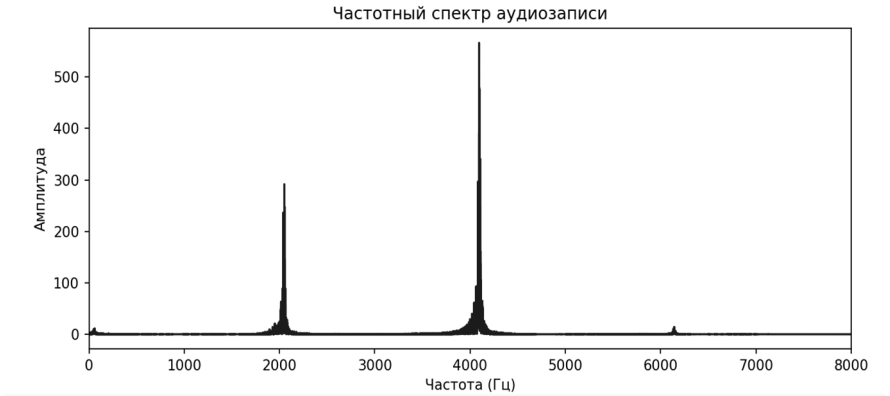

Рис. 7. Запись ускоренная, частотный спектр

Fig. 7. An accelerated recording, a frequency spectrum

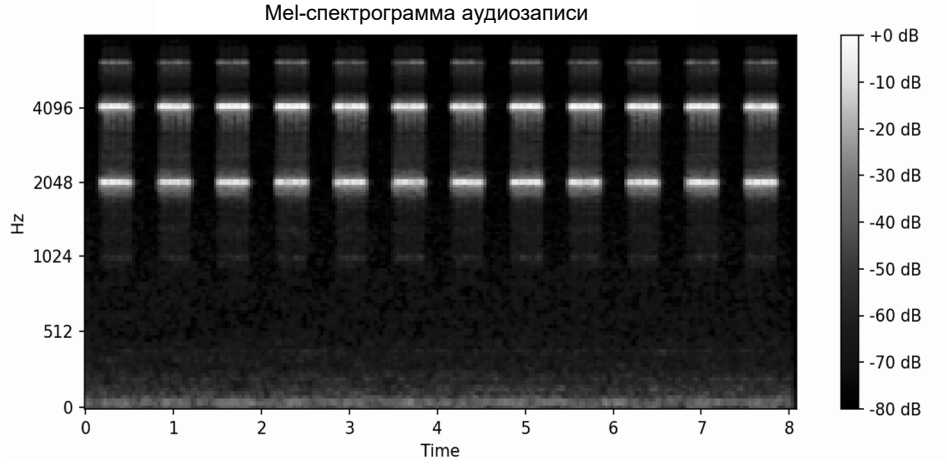

Рис. 8. Запись ускоренная, mel-спектрограмма

Fig. 8. An accelerated recording, a mel spectrogram

Эксперимент был организован следующим образом: каждому участнику (Ч1–Ч5), а также модели (М) предлагалось оценить степень сходства между оригинальной аудиозаписью 1 и некоторыми модифицированными либо иными аудиозаписями.

Полученные значения были нормированы и приведены к диапазону от нуля до единицы для дальнейшего анализа. Сравнительная табл. 2 демонстрирует степень расхождения между оценками модели и средними оценками участников:

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

где r_s – коэффициент ранговой корреляции Спирмена; n – количество наблюдений; d_i – разность рангов соответствующих значений двух переменных для i-й пары.

Для количественной оценки точности модели по отношению к мнению испытуемых был использован коэффициент ранговой корреляции Спирмена [15], рассчитываемый по формуле (1). По результатам эксперимента было рассчитано среднее значение оценок, полученных от участников, и определен коэффициент ранговой корреляции Спирмена между оценками модели и усредненными эмпирическими результатами. Полученное значение корреляции составило 0,968, что свидетельствует о высокой степени согласованности между результатами модели и восприятием человека.

Дополнительно была проанализирована точность классификации по трем основным категориям: точная копия, нечеткий дубликат и другая композиция. Наилучшая точность была достигнута в классе точных копий – 99,1 %. Для нечетких дубликатов точность составила 90 %, а для класса других композиций – 92,1 %. Это подтверждает, что модель хорошо различает даже сильно искаженные версии исходных записей, но может ошибаться в случаях, когда шумы или обрезки делают аудиофайлы пограничными по восприятию.

Таблица 2

Table 2

Сравнение результатов работы моделей и оценки человеком

Comparison of the results of the models and human evaluation

Тип испытания	Ч1	Ч2	Ч3	Ч4	Ч5	М
Аудиозапись 1, искажение	0.9	0.9	0.8	0.9	1.0	0.85
Аудиозапись 1, ускорение	0.7	0.8	0.8	0.9	0.9	0.8
Аудиозапись 1, шумы	0.9	0.5	0.7	0.7	0.8	0.68
Аудиозапись 2, совершенно другая запись	0.0	0.0	0.0	0.1	0.0	0.1
Аудиозапись 2, искажение	0.0	0.0	0.0	0.0	0.0	0.1
Аудиозапись 2, ускорение	0.0	0.0	0.0	0.0	0.0	0.05
Аудиозапись 2, шумы	0.0	0.0	0.0	0.0	0.0	0.001
Аудиозапись 3, смесь 1-й и 2-й	0.5	0.6	0.7	0.5	0.3	0.4
Аудиозапись 3, искажение	0.3	0.2	0.3	0.4	0.3	0.25
Аудиозапись 3, ускорение	0.3	0.1	0.3	0.3	0.2	0.38
Аудиозапись 3, шумы	0.2	0.3	0.2	0.2	0.3	0.21

Тем не менее модель имеет ряд ограничений. Во-первых, ее производительность может снижаться при работе с записями, содержащими перекрывающуюся речь, фоновые голоса или музыкальные вставки, не встречавшиеся в обучающем наборе.

Во-вторых, скорость обработки зависит от длины аудиозаписей, особенно при анализе больших массивов.

Таким образом, высокая корреляция Спирмена и детальная разбивка по категориям подтверждают, что предложенная модель эффективно выполняет задачу распознавания нечетких дубликатов, но нуждается в дальнейшем тестировании на более разнообразных аудиоданных.

ЗАКЛЮЧЕНИЕ

Использование различных методов сравнения аудиодубликатов дает различную точность и скорость работы. Для повышения эффективности необходимо правильно подобрать инструменты анализа. Одним из ключевых компонентов в задаче поиска нечетких дубликатов аудиозаписей является способ преобразования аудиосигнала. Преобразование в mel-спектрограмму позволяет сохранить наиболее значимые признаки, на основе которых производится сравнение.

Разработка предлагаемого каскадного ансамбля моделей позволяет значительно сократить время, необходимое для анализа больших объемов аудиоданных. Это достигается за счет использования легковесных архитектур, обеспечивающих высокую скорость обработки без существенной потери точности. Кроме того, ансамбль демонстрирует высокую точность в задаче распознавания нечетких дубликатов, что делает его особенно перспективным для практического применения.

Важно отметить, что разработанная архитектура может быть адаптирована не только для работы с аудиофайлами, но и для сравнения видеофрагментов и изображений, что существенно расширяет сферу ее применения.

СПИСОК ЛИТЕРАТУРЫ

1. Copy and move detection in audio recordings using dynamic time warping algorithm / K. Mannepalli, P. Krishna, K. Krishna, K. rama Krishna // International Journal of Innovative Technology and Exploring Engineering. – 2019. – Vol. 9. – P. 2244–2249. – DOI: 10.35940/ijitee.B6678.129219.

2. *Маленко С.А.* Увеличение производительности алгоритмов поиска дубликатов аудиозаписей // Молодой ученый. – 2017. – № 49 (183). – С. 22–26. – URL: https://moluch.ru/archive/183/47026/ (дата обращения: 08.09.2025).

3. *Ryynanen M., Klapuri A.* Query by humming of MIDI and audio using locality sensitive hashing // Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing. – IEEE, 2008. – P. 2249–2252. – DOI: 10.1109/ICASSP.2008.4518093.

4. *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://cyberleninka.ru/article/n/analiz-metodov-raspoznavaniya-i-preobrazovaniya-audioinformatsii-v-noty (дата обращения: 08.09.2025).

5. *Новохрестова Д.И.* Временная нормализация слогов алгоритмом динамической трансформации временной шкалы при оценке качества произнесения слогов // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2017. – Т. 20, № 4. – С. 142–145. – DOI: 10.21293/1818-0442-2017-20-4-142-145.

6. *Wang Y., Lyu X. Yang S.* Ocean observing time-series anomaly detection based on DTW-TRSAX method // The Journal of Supercomputing. – 2024. – Vol. 80. – P. 18679–18704. – DOI: 10.1007/s11227-024-06183-w.

7. *Ustubioglu A., Ustubioglu B., Ulutas G.* Mel spectrogram-based audio forgery detection using CNN // Signal, Image and Video Processing. – 2023. – Vol. 17. – P. 2211–2219. – DOI: 10.1007/s11760-022-02436-4.

8. 1D-CNN-based audio tampering detection using ENF signals / H. Zhao, Y. Ye, X. Shen, L. Liu // Scientific Reports. – 2024. – Vol. 14. – P. 11186. – DOI: 10.1038/s41598-024-60813-0.

9. *Wang W., Lu Z.* Few-shot bronze vessel classification via siamese fourier networks // Scientific Reports. – 2024. – Vol. 14. – P. 18011. – DOI: 10.1038/s41598-024-69272-z.

10. *Lin Y.B., Bertasius G.* Siamese vision transformers are scalable audio-visual learners // Lecture Notes in Computer Science. – 2025. – Vol. 15072. – P. 303–321. – DOI: 10.1007/978-3-031-72630-9_18.

11. *Tzanetakis G., Cook P.* Musical genre classification of audio signals // IEEE Transactions on Speech and Audio Processing. – 2002. – Vol. 10 (5). – P. 293–302. – DOI: 10.1109/TSA.2002.800560.

12. CNN architectures for large-scale audio classification / S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, K. Wilson // arXiv. – 2017. – URL: https://arxiv.org/abs/1609.09430 (accessed: 08.09.2025).

13. *Ананьев А.С., Бутенко Д.В., Попов К.В.* 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: 08.09.2025).

14. *Бурцев А.Г., Мельников А.В.* Численное моделирование и анализ спектра системы прерывающихся сигналов // Инженерный вестник Дона. – 2014. – № 2. – URL: http://www.ivdon.ru/ru/magazine/archive/n2y2014/2314 (дата обращения: 08.09.2025).

15. *Кошелева Н.Н.* Корреляционный анализ и его применение для подсчета ранговой корреляции Спирмена // Актуальные проблемы гуманитарных и естественных наук. – 2012. – № 5. – С. 23–26. – URL: https://cyberleninka.ru/article/n/korrelyatsionnyy-analiz-i-ego-primenenie-dlya-podscheta-rangovoy-korrelyatsii-spirmena (дата обращения: 08.09.2025).

Левшин Денис Витальевич, магистрант Волгоградского государственного технического университета. E-mail: levshin01@bk.ru

Скляров Михаил Александрович, магистрант Волгоградского государственного технического университета. E-mail: mikhail.29.06.2001@gmail.com

Быстряков Даниил Владимирович, магистрант Волгоградского государственного технического университета. E-mail: bystriackoff@yandex.ru

Зубков Александр Владимирович, кандидат технических наук, доцент кафедры программного обеспечения автоматизированных систем Волгоградского государственного технического университета. Основное направление исследований – методы автоматизированного анализа гетерогенных семантических данных информационных систем. Имеет более 130 научных работ, в том числе 3 учебных пособия. E-mail: aleksandr.zubkov@volgmed.ru

Levshin Denis V., Master's student at Volgograd State Technical University. E-mail: levshin01@bk.ru

Sklyarov Mikhail A., Master's student at Volgograd State Technical University. E-mail: mikhail.29.06.2001@gmail.com

Bystryakov Daniil V., Master's student at Volgograd State Technical University. E-mail: bystriackoff@yandex.ru

Zubkov Alexander V., PhD (Eng.), associate Pprofessor, Department of Software for Automated Systems, Volgograd State Technical University. The main research area is methods of automated analysis of heterogeneous semantic data of information systems. He has more than 130 scientific papers, including 3 textbooks. E-mail: aleksandr.zubkov@volgmed.ru

DOI: 10.17212/2782-2001-2025-3-69-82

The use of convolutional and siamese neural networks in the task of comparing fuzzy duplicates of audio objects[*]

D.V. LEVSHIN[a], D.V. BYSTRYAKOV[b], M.A. SKLYAROV[c], A.V. ZUBKOV[d]

Volgograd State Technical University, 28 Lenin Avenue, Volgograd, 400005, Russian Federation

[a] levshin01@bk.ru [b] bystriackoff@yandex.ru [c] mikhail.29.06.2001@gmail.com [d]aleksandr.zubkov@volgmed.ru

Abstract

At present, there is a growing volume of data in the form of audio recordings, which are difficult to process due to a large number of duplicates, and noisy, or truncated recordings. This paper examines one possible solution to the problem of identifying fuzzy duplicates of audio

[*] *Received 17 February 2025.*

recordings in large datasets. The proposed solution is based on the use of a cascade ensemble of models to detect fuzzy duplicates. To extract features, analyze temporal parameters, and assess similarity between recordings, Convolutional Neural Networks (CNN), Temporal Shift Networks (TSN), and Siamese neural networks were employed. The audio data input into the models was first transformed into a set of spectrograms using the Short-Time Fourier Transform (STFT) algorithm. Each audio recording was segmented at a specified sampling rate, transformed using STFT, and then passed to the ensemble of models. The study focuses on the behavior of the ensemble when processing recordings subjected to various alterations such as noise, distortion, or truncation. Experiments conducted on the dataset demonstrated a high degree of correlation between the results obtained by a group of human participants and those produced by the model ensemble, confirming the effectiveness of the proposed approach. The ensemble showed strong robustness to different types of audio modifications, including tempo changes, added noise, and cropping. Future research may focus on adapting the ensemble for different types of data, including video and graphical content, thereby expanding the applicability of the proposed method.

Keywords: fuzzy duplicate, convolution, neural network, STFT, noise, performance, spectrogram, Siamese network, segmentation, cascaded ensemble

REFERENCES

1. Mannepalli K., Krishna P., Krishna K., Krishna K. rama. Copy and move detection in audio recordings using dynamic time warping algorithm. *International Journal of Innovative Technology and Exploring Engineering*, 2019, vol. 9, pp. 2244–2249. DOI: 10.35940/ijitee.B6678.129219.

2. Malenko S.A. Uvelichenie proizvoditel'nosti algoritmov poiska dublikatov audiozapisei [Increasing the performance of duplicate audio search algorithms]. *Molodoi uchenyi = Young Scientist*, 2017, no. 49 (183), pp. 22–26. Available at: https://moluch.ru/archive/183/47026/ (accessed: 08.09.2025).

3. Ryynanen M., Klapuri A. Query by humming of MIDI and audio using locality sensitive hashing. *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 2249–2252. DOI: 10.1109/ICASSP.2008.4518093.

4. Bulavin D.A., Kharitonov I.A. Analiz metodov raspoznavaniya i preobrazovaniya audioinformatsii v noty [Analysis of audio recognition and conversion methods to musical notes]. *Avtomatizirovannye sistemy upravleniya i pribory avtomatiki = Management Information System and Devices*, 2011, no. 152 (2), pp. 56–63. Available at: https://cyberleninka.ru/article/n/analiz-metodov-raspoznavaniya-i-preobrazovaniya-audioinformatsii-v-noty (accessed 08.09.2025).

5. Novokhrestova D.I. Vremennaya normalizatsiya slogov algoritmom dinamicheskoi transformatsii vremennoi shkaly pri otsenke kachestva proizneseniya slogov [Time normalization of syllables with the dynamic time warping algorithm in assessing of syllables pronunciation quality when speaking]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki = Proceedings of TUSUR University*, 2017, vol. 20, no. 4, pp. 142–145. DOI: 10.21293/1818-0442-2017-20-4-142-145.

6. Wang Y., Lyu X. Yang S. Ocean observing time-series anomaly detection based on DTW-TRSAX method. *The Journal of Supercomputing*, 2024, vol. 80, pp. 18679–18704. DOI: 10.1007/s11227-024-06183-w.

7. Ustubioglu A., Ustubioglu B., Ulutas G. Mel spectrogram-based audio forgery detection using CNN. *Signal, Image and Video Processing*, 2023, vol. 17, pp. 2211–2219. DOI: 10.1007/s11760-022-02436-4.

8. Zhao H., Ye Y., Shen X., Liu L. 1D-CNN-based audio tampering detection using ENF signals. *Scientific Reports*, 2024, vol. 14, p. 11186. DOI: 10.1038/s41598-024-60813-0.

9. Wang W., Lu Z. Few-shot bronze vessel classification via siamese Fourier networks. *Scientific Reports*, 2024, vol. 14, p. 18011. DOI: 10.1038/s41598-024-69272-z.

10. Lin Y.B., Bertasius G. Siamese vision transformers are scalable audio-visual learners. *Lecture Notes in Computer Science*, 2025, vol. 15072, pp. 303–321. DOI: 10.1007/978-3-031-72630-9_18.

11. Tzanetakis G., Cook P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 2002, vol. 10 (5), pp. 293–302. DOI: 10.1109/TSA.2002.800560.

12. 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*. 2017. Available at: https://arxiv.org/abs/1609.09430 (accessed 08.09.2025).

13. Ananyev A.S., Butenko D.V., Popov K.V. Modelirovanie protsessov upravleniya kachestvom produktsii na osnove imitatsionnogo modelirovaniya [Intelligent technologies for design information systems. The method of design software products in the presence of a prototype]. *Inzhenernyi vestnik Dona = Engineering Journal of Don*, 2012, no. 2. Available at: http://www.ivdon.ru/ru/magazine/archive/n2y2012/815 (accessed 08.09.2025).

14. Burtsev A.G., Melnikov A.V. Chislennoe modelirovanie i analiz spektra sistemy preryvayushchikhsya signalov [Numerical modeling and analyze of discontinuous signals spectrumenterprise]. *Inzhenernyi vestnik Dona = Engineering Journal of Don*, 2014, no. 2. Available at: http://www.ivdon.ru/ru/magazine/archive/n2y2014/2314 (accessed 08.09.2025).

15. Kosheleva N.N. Korrelyatsionnyi analiz i ego primenenie dlya podscheta rangovoi korrelyatsii Spirmena [Correlation analysis and its application for calculating Spearman's rank correlation]. *Aktual'nye problemy gumanitarnykh i estestvennykh nauk = Actual Problems of Humanities and Natural Sciences*, 2012, no. 5, pp. 23–26. Available at: https://cyberleninka.ru/article/n/korrelyatsionnyy-analiz-i-ego-primenenie-dlya-podscheta-rangovoy-korrelyatsii-spirmena (accessed 08.09.2025).

Для цитирования:

Применение сверточных и сиамских нейронных сетей в задаче сравнения нечетких дубликатов аудиообъектов / Д.В. Левшин, Д.В. Быстряков, М.А. Скляров, А.В. Зубков // Системы анализа и обработки данных. – 2025. – № 3 (99). – С. 69–82. – DOI: 10.17212/2782-2001-2025-3-69-82.

For citation:

Levshin D.V., Bystryakov D.V., Sklyarov M.A., Zubkov A.V. Primenenie svertochnykh i siamskikh neironnykh setei v zadache sravneniya nechetkikh dublikatov audioob"ektov [The use of convolutional and siamese neural networks in the task of comparing fuzzy duplicates of audio objects]. *Sistemy analiza i obrabotki dannykh = Data Analysis and Processing Systems*, 2025, no. 3 (99), pp. 69–82. DOI: 10.17212/2782-2001-2025-3-69-82.