Научная статья

УДК 004.032.26
DOI 10.25205/1818-7900-2022-20-3-38-50

Метод для восстановления аудиосигнала с помощью свёрточных нейронных сетей

Кристина Игоревна Дементьева[1], Антон Андреевич Ракитский[2]

[1, 2]Сибирский государственный университет телекоммуникаций и информатики
Новосибирск, Россия

[2]Новосибирский государственный университет
Новосибирск, Россия

[1]k.morozova@cyber.sibsutis.ru, https://orcid.org/ 0000-0001-8650-3749
[2]rakitsky.anton@gmail.com, https://orcid.org/0000-0002-1224-4278

Аннотация

В работе представлено исследование возможности восстановления искажённого аудио-сигнала. На основе полученных ранее результатов использования методов глубокого машинного обучения разработана концепция нейронной сети, предназначенная для коррекции искаженного звукового сигнала. На базе первоначально полученной разработано несколько новых архитектур нейронных сетей, ориентированных на восстановление аудиосигнала. В статье приведены описания разработанных архитектур с теоретическим обоснованием возможности их применения. Представленные архитектуры были протестированы для решения задачи восстановления партии конкретного инструмента в музыкальной композиции, из которой она была удалена.

Ключевые слова

аудиосигнал, восстановление сигнала, методы машинного обучения, нейронные сети, регрессия

Для цитирования

Дементьева К. И., Ракитский А. А. Метод для восстановления аудиосигнала с помощью свёрточных нейронных сетей // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 38–50. DOI 10.25205/1818-7900-2022-20-3-38-50

Method for Restoring Audio Signal Using Convolutional Neural Networks

Kristina. I. Dementyeva[1], Anton A. Rakitskiy[2]

[1, 2]Siberian State University of Telecommunications and Information Sciences
Novosibirsk, Russian Federation

[2]Novosibirsk State University
Novosibirsk, Russian Federation

[1]k.morozova@cyber.sibsutis.ru, https://orcid.org/ 0000-0001-8650-3749
[2]rakitsky.anton@gmail.com, https://orcid.org/0000-0002-1224-4278

Abstract

This paper describes a research of the restoring distorted audio signal possibility. Based on the previously obtained results of using deep machine learning methods, the concept of a neural network to correct a distorted audio signal has been developed. On the basis of the originally obtained results, several new neural network architectures were developed, focused on the audio signal restoring. The paper contains descriptions of the developed architectures with a theoretical substantiation of the possibility of their application. The presented architectures were tested to solve the problem of restoring the part of a specific instrument in a musical composition where it was removed. The results of testing the developed architectures of neural networks are presented in several forms.

Keywords
audio signal, signal recovery, machine learning methods, neural networks, regression

For citation
Dementyeva K. I., Rakitsky A. A. Method for Restoring Audio Signal Using Convolutional Neural Networks. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 38–50. DOI 10.25205/1818-7900-2022-20-3-38-50

Введение

Среди областей обработки сигналов можно выделить такие, как: наложение эффектов, распознавание и устранение шума, усиление сигнала, улучшение качества и частотных характеристик, восстановление в случае потерь, восстановление удаленной части сигнала и др. Восстановление направлено на исправление искаженной части таким образом, чтобы она была наиболее близкой к оригиналу.

Искажения разного рода могут возникать в аудиосигнале при его формировании, передаче, записи или обработке. Например, при записи музыкальных партий в студии возможно возникновение сбоев в работе звукозаписывающего оборудования, что приводит к потере части аудиосигнала. Не исключено, что сходящий с конвейера носитель информации аудио-сигнала имеет дефекты, которые приведут к повреждению сигнала. При распространении звуковой волны в какой-либо среде также могут возникнуть искажения (например, телефония, эхолокация и т. д.), вызванные как особенностями самой среды, так и внешним воздействием. В процессе эксплуатации носителя аудиосигнала также могут появиться новые повреждения, приводящие к потере данных. Существуют и другие возможности повреждения аудиосигнала. В данной работе предлагается новый способ исправления таких искажений, базирующийся на свёрточных нейронных сетях.

В последующих главах формулируется модель искажения, которая наиболее адаптирована для применения к ней нейронных сетей, описан процесс исследования структуры аудиосигнала, рассмотрены различные архитектуры нейронных сетей и представлена архитектура, которая позволяет решить задачу восстановления сигнала с высокой точностью. В общем случае, исследование конкретных практических задач восстановления сигнала влияет на потенциальную возможность дальнейшей разработки инновационного программного обеспечения, позволяющего доносить до пользователя истинную форму сигнала. В разделах «Особенности реализации» и «Анализ результатов» разработанные архитектуры исследуются в рамках применения к задаче восстановления ударной партии в музыкальной композиции, из которой она была удалена.

1. Анализ состояния предметной области

В настоящее время известен ряд методов восстановления сигналов, в частности: в работе [1] исследована задача восстановления аудио на основе заполнения недостающих сегментов аудио. Исследователи в рамках данной работы предлагают улучшения в наложении звука, направленные на компенсацию потери в аудиосигнале.

В исследовании [2] предложены методы восстановления аудиофайлов, основанные на нейронных сетях с прямой связью и долговременной памятью, а также описаны возможности автоматизации задачи восстановления сигнала с помощью предлагаемых методов восстановления, использующих глубокие нейронные сети.

Подходы к идентификации и восстановлению структурированных данных на основе методов глубокого обучения описаны в [3–5]. Реконструкция звуковых сигналов, искаженных щелчками и потрескиваниями, методами сжатия и восстановления представлена в [6].

В исследовательской статье [7] предлагается алгоритм на основе многослойной сверточной нейронной сети для улучшения звукового сигнала. Частотные характеристики звукового сигнала были извлечены как набор частот, которые кратко описывают общую форму спектраль-

ной огибающей. Поскольку аудиосигнал изменяется с очень постоянной скоростью, предполагается, что аудиосигнал не сильно изменяется за короткие промежутки времени. Предлагаемая нейронная сеть принимает входные данные от разных кадров из звукового сигнала, загрязненного шумом, для обучения и тестирования.

В отличие от описанных выше методов, в данной работе исследуется возможность восстановления звукового сигнала с помощью регрессионного анализа. В частности, рассматривается проблема восстановления партии ударных в музыкальном произведении, где она была удалена, и способы ее решения на основе методов машинного обучения.

2. Описание модели искажения и восстановления аудиосигнала

Восстановление сигнала – реконструкция истинной формы сигнала, несущего полезное сообщение после его искажения зашумлением либо удалением какой-либо его части.

Предположим, что при восстановлении аудиосигнала можно следовать следующей схеме: используя кратковременное преобразование Фурье для получения таких характеристик сигнала, как синусоидальная частота и фазовое содержание локальных участков сигнала по мере его изменения во времени и, разделив звуковой сигнал на небольшие промежутки времени, можно представить его в виде «изображения». Благодаря произведенным действиям, «изображения» сигнала имеют пространственные закономерности во временной и частотной области, что позволяет обучить искусственный интеллект обнаруживать отсутствие составляющих аудиосигнала на конкретном частотном интервале в конкретный момент времени.

В исходном сигнале покадрово выполняется преобразование Фурье и, таким образом, определяется частотный спектр каждого кадра аудиосигнала. Опишем это преобразование подробно: каждая часть исходного сигнала подвергается преобразованию Фурье для получения m частотных диапазонов компонент сигнала из рассматриваемого кадра; затем отсекается половина полученного спектра, поскольку для входного сигнала с действительным знаком спектр в диапазоне от $1/2T$ до $1/T$ просто отражает спектр от 0 до $1/2T$ (где $1/T$ – частота дискретизации) и, соответственно, не несет полезной нагрузки.

Таким образом, получен набор кадров, который представляет информацию о значениях всех частотных диапазонов в конкретных временных интервалах. В каждом конкретном кадре необходимо восстановить недостающую часть ударника или выявить, что она там не нужна; делать это нужно с учетом зависимости от соседних кадров.

Рассмотрим некоторый аудиосигнал длительностью t секунд, записанный с частотой дискретизации R, в котором необходимо восстановить какую-то его составляющую (например, в музыкальной композиции была вырезана партия ударника). Разделим этот сигнал на n равных временных интервалов (кадров), которые будут содержать m значений сигнала. К каждому кадру рассматриваемого сигнала применяется преобразование Фурье.

Пусть $X_1, X_2, \ldots, X_n$ – векторы со значениями исходного сигнала, разделенными на кадры, а функция $f(X)$ – преобразование Фурье кадра X. Тогда $f(X_1), f(X_2), \ldots, f(X_n)$ – это комплексные векторы, полученные после преобразования Фурье каждого из кадров исходного сигнала. Можно представить совокупность преобразований всех кадров в виде матрицы комплексных чисел $F = f_1, \ldots, f_n$ с размерностью (n, m), где $f_i = f(X_i)$.

Обозначим $\bar{F}$ как матрицу комплексных чисел размерностью (n, m), которая является дополнением к матрице F, необходимой для восстановления сигнала. Тогда $F_r = F + \bar{F}$, где F_r – это матрица значений преобразования Фурье для каждого кадра восстановленного сигнала. Чтобы получить сам восстановленный аудиосигнал, к F_r следует применить обратное преобразование Фурье.

Предполагается, что $\bar{F}$ имеет некоторую функциональную зависимость от F. Опишем функцию зависимости дополнения к F от исходного искаженного F как Z: $F \rightarrow \bar{F}$. В рамках данной работы будем предполагать, что такая зависимость Z существует. Тогда решение задачи

восстановления звукового сигнала F_t сводится к выявлению зависимости Z. Исходя из вышесказанного, можно попытаться выявить эту сложную функциональную зависимость с помощью нейронных сетей.

Переходя к конкретной задаче, предположим, что k-й частотный пул t-го кадра $a_{k,t}$ зависит от всех других частотных пулов всех кадров,

$$a_{k,t} = \sum_{i=0}^{n} \sum_{j=0}^{m} G_{i,j,k,t}(a_{i,j}) + \varepsilon,$$

где $G_{i,j,k,t}$ – некоторая неизвестная нелинейная функция, описывающая влияние $a_{i,j}$ на $a_{k,t}$, а ε – случайная составляющая (это значение должно быть довольно маленьким и описывать некоторые недостаточные отклонения в представленной зависимости). Поскольку функции $G_{i,j,k,t}$ неизвестны, их логично аппроксимировать с помощью нейронных сетей.

3. Исследование особенностей восстановления аудиосигнала

Работа с исследованием аудиосигнала имеет ряд особенностей: задачу восстановления аудиосигнала следует рассматривать как задачу линейной регрессии комплексных чисел, размер исходных файлов достаточно велик для быстрой обработки, следует рассматривать каждый канал звукового файла отдельно для более точной обработки.

3.1. Задача линейной регрессии комплексных чисел

При записи звукового сигнала с помощью микрофона, он улавливается и преобразуется в электрический сигнал. Затем преобразованный сигнал посылается звуковой карте на компьютере, которая преобразует сигнал в числа, эти числа называются сэмплами. Сэмплы представляют собой информацию о том, как записанный сигнал звучит в определённые моменты времени. Например, для создания цифрового аудиосигнала, качеством схожим с обычной записью на компакт-диске, должна поступать информация о 44 100 сэмплах в секунду. Количество сэмплов, полученных в секунду, называется частотой сэмплирования или частотой дискретизации. Размер каждого отдельного сэмпла в свою очередь тоже влияет на качество записываемого звука, этот размер называется разрешением. Чем больше разрешение, тем выше качество звука. Для цифровой звукозаписи с таким же качеством, как на компакт-диске, каждый сэмпл будет размером 16 бит.

После считывания аудиосигнала данные представляют собой числа типа int16. Для получения более конкретных характеристик для последующей обработки сигнала необходимо знать фазу и амплитуду сигнала на конкретном частотном интервале в конкретный период времени, для этого необходимо, чтобы сигнал подвергся преобразованию Фурье. Данное преобразование возвращает аудиосигнал, представляющий собой данные типа complex128, амплитуда сигнала определяется как модуль комплексного числа, фаза – как аргумент.

Таким образом, задача восстановления сигнала сводится к следующей: имеются числовые данные в виде комплексных чисел, нужно восстановить число, показывающее характеристику сигнала на заданной частоте в определенный момент времени. То есть задача линейной регрессии в данном случае представляет собой линейную регрессию комплексных чисел.

3.2. Объем входных данных

Исходные данные представлены в формате WAV (WAVE, WAV, от англ. waveform – «в форме волны»), использование данного формата аудиосигнала позволяет использовать данные в несжатом виде и без потерь, в то время как более привычный для пользователя формат

MP 3 содержит сжатые данные с потерями, что позволяет хранить ту же самую аудиоинформацию в меньшем размере файла.

Использование данных в несжатом формате подразумевает работу с большим объёмом данных. Каждый аудиосигнал имеет следующие характеристики: количество каналов, частота дискретизации (показывает, сколько раз в секунду производится выборка аудио), разрешение (количество бит, используемых для кодирования каждого значения выборки) и продолжительность. Для хранения одной многоканальной стереозаписи длительностью 3 минуты с частотой дискретизации 44,1 кГц и разрешением 16 бит потребуется порядка 30 Мбайт. Для качественного обучения моделей глубокого машинного обучения могут потребоваться десятки и сотни аудиозаписей, что при сложных архитектурах моделей будет означать большое количество времени, требуемое для обработки информации.

3.3. Многоканальная аудиоинформация

Зачастую используется стерео аудиосигнал. Стереофонический звук состоит из двух уникальных каналов. При прослушивании это позволяет получить объемное звучание, при котором есть ощущение направленности звука и его расположения. На практике же стереозвук представляет собой сигнал, что для записи которого используются два микрофона, каждый из которых улавливает немного отличающийся сигнал.

Некоторые кодеки фактически разделяют левый и правый каналы, сохраняя их в отдельных блоках в своей структуре данных. Таким образом, при исследовании стереосигнала мы имеем два моносигнала с небольшими различиями. Это приводит к тому, что, если необходимо получить более точное предсказание стереосигнала, нужно рассматривать и обрабатывать обе его составляющие отдельно.

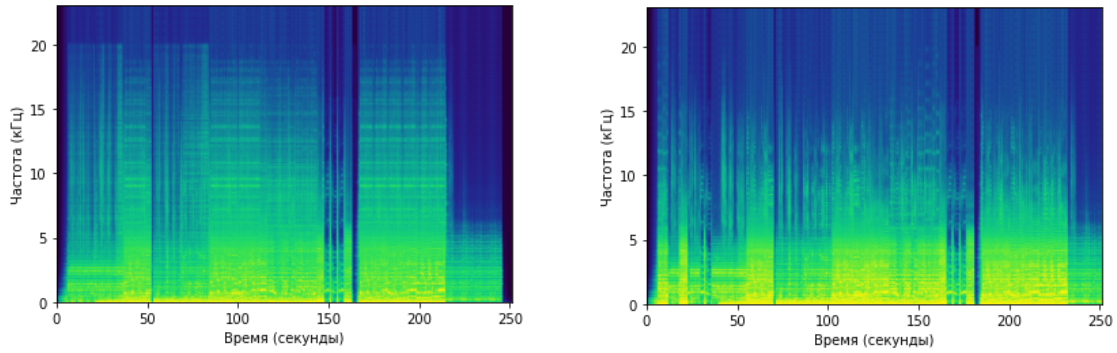
4. Особенности реализации

Частная задача восстановления сигнала – восстановление партии ударника в аудиосигнале (музыкальном произведении). Актуальность и сложность исследования заключается в том, что частотные диапазоны ударных на фоне всего частотного разложения сигнала трудно определить, поэтому процесс восстановления должен основываться на всем исходном частотном диапазоне.

Создание партии ударных – это уникальный процесс, в котором человек придумывает ритмический рисунок песни, но он так или иначе будет соответствовать некоторым ритмическим законам и будет зависеть от других инструментов (или, наоборот, партии инструментов будут зависеть от партии ударника). Таким образом, весь частотный диапазон аудиосигнала несет в себе значимые признаки для последующего восстановления барабанной партии, которая является дополнением к исходному сигналу без нее.

Исходные данные в данной работе – аудиосигнал с удаленной барабанной партией и аудиосигнал с присутствующей партией.

Различия оригинального аудиосигнала от того же аудиосигнала с удаленной барабанной партией в частотных интервалах можно наблюдать после применения к каждому сигналу разложения Фурье. На рис. 1а представлена спектрограмма композиции исполнителя 30 Seconds to Mars, на рис. 1б – спектрограмма той же композиции с удаленной барабанной партией. Можно наблюдать, что в первом случае при наличии барабанной партии некоторые изображения частотных интервалов более выраженные, яркие, полный частотный спектр аудиосигнала выглядит более дополненным, чем второй случай с удалённой барабанной партией.

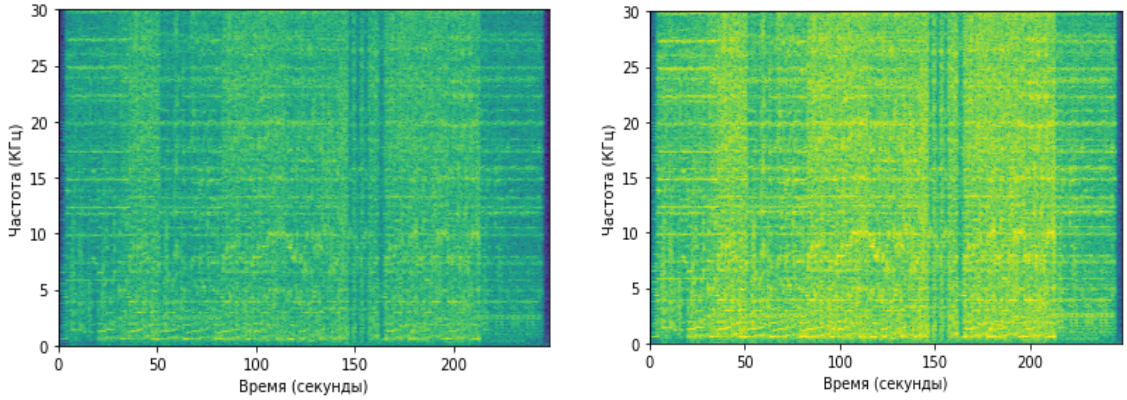


(а) (б)

Рис. 1. Спектрограмма композиции исполнителя 30 Seconds to Mars оригинал (а) и с удаленной барабанной партией (б)
Fig. 1. Spectrogram of the composition of the artist 30 Seconds to Mars original (a) and without a drum part (б)

Изначально аудиосигналы поступают на обработку в формате WAV. После считывания в данном формате данные представляют собой массив чисел типа int16. Затем необходимо разделить исходный стереосигнал на два моносигнала, чтобы обеспечить детальную обработку каждой составляющей сигнала отдельно, и произвести преобразование Фурье для каждого канала сигнала. Во время преобразования следует обратиться к нормализации данных.

Нормализация данных – это операция преобразования входной информации (признаков), которая выполняется на этапе подготовки данных (при генерации признаков) в машинном обучении, при которой значения признаков приводятся к некоторому заданному диапазону. После нормализации все числовые значения входных признаков приведены к одинаковой области их изменения – некоторому узкому диапазону, что обеспечит корректную работу вычислительных алгоритмов. На рис. 2 показаны частотные спектры аудио-сигнала после преобразования Фурье, при создании которого использовалась (а) и не использовалась нормализации (б).



(а) (б)

Рис 2. Спектрограмма аудиосигнала после преобразования Фурье, при создании которого использовалась (а) и не использовалась нормализации (б)
Fig. 2. Spectrogram of the audio signal after the Fourier transform, which was created using (a) and not using normalization (б)

После преобразований Фурье данные представляют собой многомерный массив размером (512, *n*, 2), где *n* – количество временных интервалов (кадров) аудиосигнала, а 2 – показатель того, что комплексные числа рассматриваются как реальная и мнимая часть отдельным значением.

Задачу восстановления удаленной части аудиосигнала предлагается решать с помощью свёрточной нейронной сети. Данный тип методов глубокого обучения является одним их самых популярных на данный момент, а активное применение получил, начиная с 2012 года. Главное преимущество свёрточных нейронных сетей состоит в том, что они автоматически извлекают нужные признаки и коэффициенты без какого-либо человеческого наблюдения. Также эффективны в вычислениях, в них используются специальные операции свертки и объединения, а также выполняются объединенное использование параметров.

Для восстановления одного кадра рассматриваются 25 последовательных кадров (предположим, что этого количества достаточно, чтобы установить все сложные зависимости во времени и между разными частотными диапазонами). Итак, из 25 кадров восстанавливаем 1 кадр посередине. Первоначальное решение заключалось в восстановлении всех 512 частотных диапазонов входного аудиосигнала с помощью одной нейронной сети. В данном случае использовалась архитектура свёрточной нейронной сети, которая восстанавливает (за счет реализованного скользящего окна с шагом в один временной интервал) весь частотный диапазон на одном конкретном временном интервале, работая с размерами (512, 25, 2) на входе и (512, 1, 2) на выходе.

Данная архитектура сети имеет около 336 миллионов параметров, поэтому как обучить, так и работать с такой сетью – довольно трудоемкая задача. Тем не менее, получить качественные результаты не удалось, поскольку сеть создавала на выходе зашумленную версию исходного аудиосигнала. Для преодоления проблемы было решено изменить конструкцию нейронной сети, добавив на выходе слой, содержащий исходный сигнал, соответствующий результирующему кадру. Также было решено сократить количество используемых параметров до ~ 4 миллионов. Оказалось, что результаты становятся близкими к предыдущему эксперименту – сеть генерировала зашумленный исходный сигнал без ожидаемых изменений.

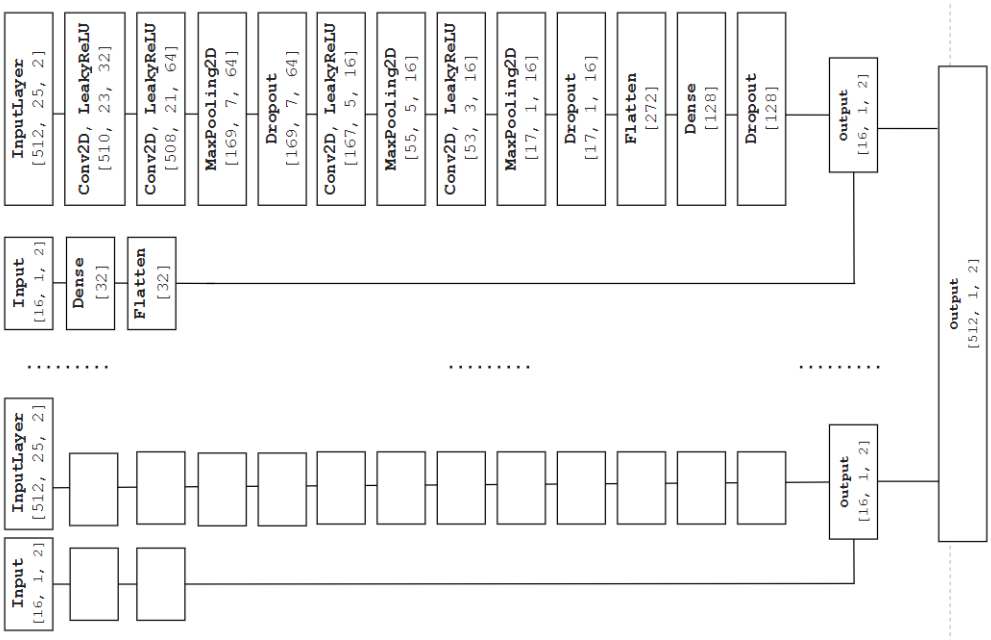


Рис 3. Схема объединения нейронных сетей, разработанных на непересекающихся частотных поддиапазонах
Fig. 3. Scheme for combining neural networks developed on non-overlapping frequency subranges

После получения представленных выше результатов архитектура была изменена таким образом, чтобы количество параметров было уменьшено до 70 тысяч (см. рис.3). Кроме того,

внесены следующие изменения: теперь весь частотный диапазон из 512 полос делится на части, и каждая эта часть обрабатывается отдельной нейронной сетью. Таким образом, каждая нейронная сеть обрабатывает только свою часть всего частотного диапазона. Предполагалось, что такой подход должен повысить точность восстановления. Если количество нейронных сетей равно 32, – то каждая сеть на выходе восстанавливает 16 полос из 512. Важно отметить, что на вход подаются все 512 частотных полос, но каждая сеть в данном случае восстанавливает только 16 поло. Затем результат работы всех 32 нейронных сетей складывается для получения всего частотного диапазона из 512 полос. Конечный этап – сложить полученный результат с исходным сигналом. Таким образом, выполняется не чистая модификация сигнала, а сложение исходного сигнала с восстановленными недостающими элементами.

При разработке архитектуры нейронной сети в качестве слоя активации был выбран LeakyRelu. В общем случае функция активации определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения. Входные данные, поступившие на слой, умножаются на веса полносвязного или свёрточного слоя, а результат передаётся в функцию активации. Выбранная функция LeakyRelu имеет преимущество относительно большинства подобных функций активации – отрицательные значения обрабатываются благодаря небольшому наклону в левой полуплоскости. Математически LeakyReLu объявлена следующим образом:

$$f(x) = \begin{cases} 0.01x, x < 0 \\ x, x \geq 0 \end{cases}$$

В качестве функции потерь применена функция среднеквадратичного отклонения. Функция потерь в теории статистических решений характеризует потери при неправильном принятии решений на основе наблюдаемых данных. Другими словами, использование функции потерь позволяет получить оценку того, насколько хорошо алгоритм моделирует набор данных – если прогнозы полностью неверны, функция потерь выдаст большее число, если они достаточно хороши – будет выведено меньшее число. Чтобы вычислить среднеквадратичную ошибку необходимо вычислить разницу между предсказанными и исходными данными, возвести ее в квадрат и усреднить по всему набору данных. Цель использования функции потерь – отследить и минимизировать это значение, чтобы получить лучшую линию, проходящую через все точки при рассмотрении данных на координатной плоскости.

В процессе компилирования модели нейронной сети используется оптимизатор. Оптимизаторы – это алгоритмы или методы, используемые для изменения атрибутов нейронной сети, таких как веса и скорость обучения, с целью уменьшения потерь [8]. В рамках данного исследования использовался оптимизатор Adam [9] – адаптивный алгоритм оптимизации скорости обучения.

В качестве дополнительного метода улучшения результатов можно рассмотреть возможность увеличения количества нейронных сетей, а, следовательно, уменьшения частотного диапазона, обрабатываемого каждой сетью (например, использование 64 сетей и, соответственно, 8 диапазонов частот для обрабатываемого диапазона 512). В результате полученные частотные дополнения исходного аудиосигнала становятся более выраженными, значения восстановленного аудиосигнала приближаются к оригинальному (исходному) из-за того, что коэффициенты сети становятся настроенными на более конкретный набор неизвестных.

5. Анализ результатов

На рис. 4 и рис. 6–9 показаны спектрограммы исходного звукового сигнала (с барабанной частью) и звукового сигнала, полученного с помощью нейронной сети, соответственно. На этих рисунках можно даже визуально заметить схожесть исходного и восстановленного

сигнала, что является показателем качества метода. Кроме того, для лучшего сравнения на рис. 5 приведена аналогичная спектрограмма звукового сигнала с удаленной барабанной частью.

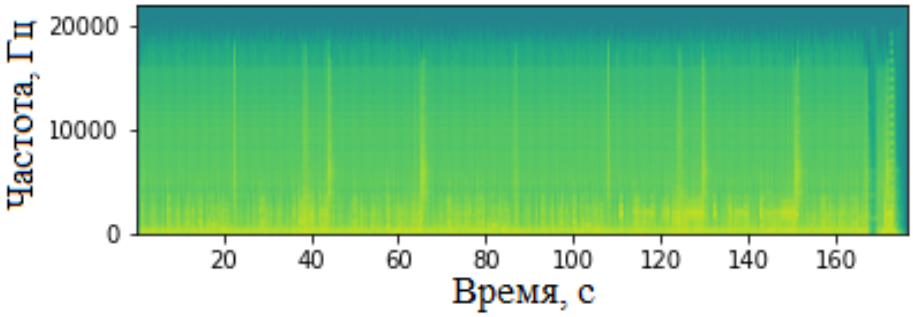


Рис 4. Спектрограмма исходного аудиосигнала с присутствующей барабанной партией
Fig. 4. Spectrogram of the original audio signal with drum part

Для улучшения результатов было увеличено количество нейронных сетей, а значит, уменьшен частотный диапазон, обрабатываемый каждой сетью (например, используя 64 сети и, соответственно, 8 частотных диапазонов из 512 для каждой сети). Полученные частотные дополнения исходного аудиосигнала становятся более выраженными, восстановленный аудиосигнал становится ближе к оригиналу из-за того, что коэффициенты сети становятся настроенными на более конкретный набор неизвестных. На рис. 6–9 показаны спектрограммы исходного аудиосигнала (с барабанной частью) и аудиосигнала, полученного с помощью нейронной сети с 16, 32, 64, 128 сетями соответственно. На этих рисунках можно даже визуально заметить схожесть исходного и восстановленного сигнала, что является показателем качества метода.

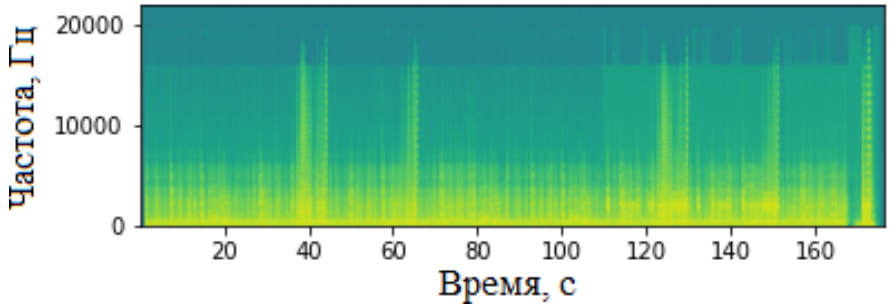


Рис 5. Спектрограмма исходного аудиосигнала с удалённой барабанной партией
Fig. 5. Spectrogram of the original audio signal without drum part

Таким образом, как видно на рис. 6–9, при увеличении количества нейронных сетей результат работы приближается к ожидаемому, но при этом увеличивается уровень шума аудиосигнала.

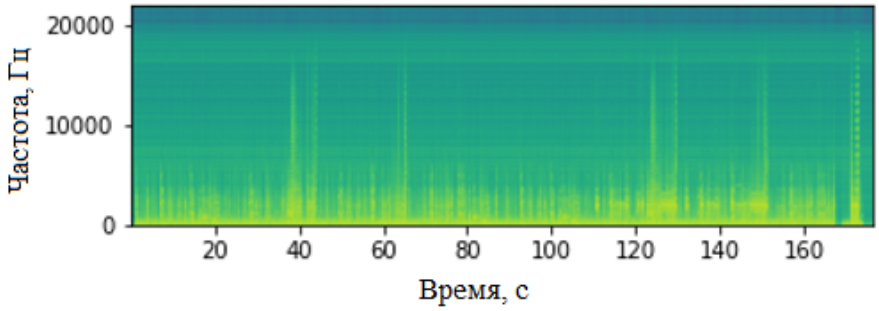


Рис 6. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 16 нейронных сетей
Fig. 6. Spectrogram of the resulting audio signal with the reconstructed drum part based on 16 neural networks

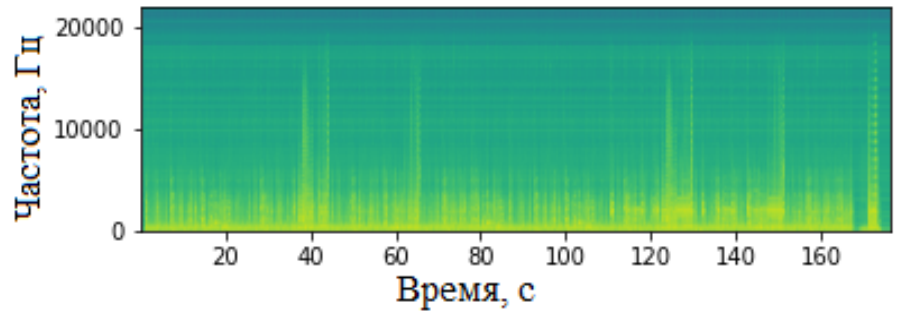


Рис 7. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 32 нейронных сетей
Fig. 7. Spectrogram of the resulting audio signal with the reconstructed drum part based on 32 neural networks

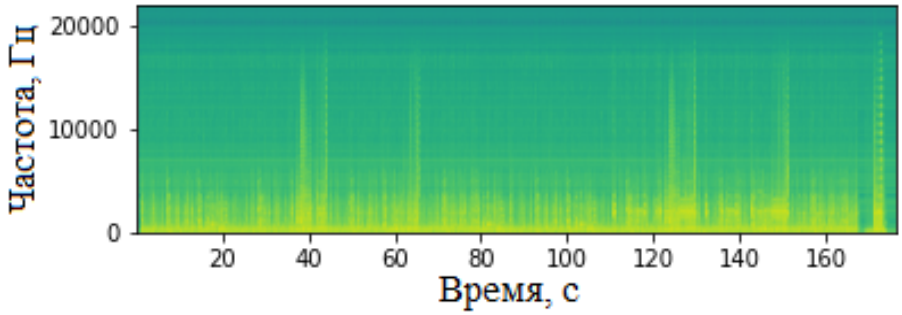


Рис 8. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 64 нейронных сетей
Fig. 8. Spectrogram of the resulting audio signal with the reconstructed drum part based on 64 neural networks

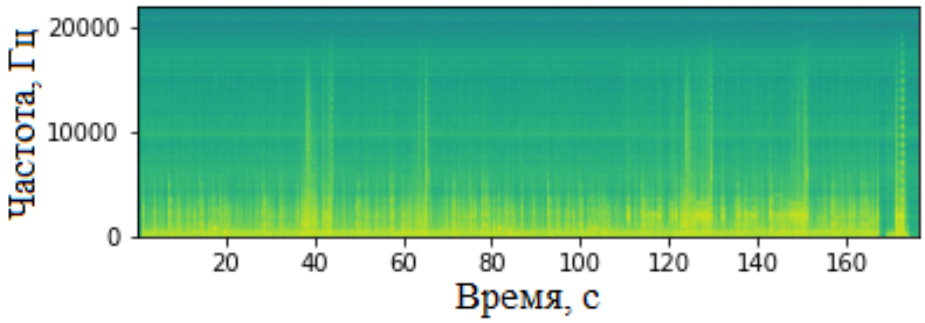


Рис 9. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 128 нейронных сетей
Fig.9. Spectrogram of the resulting audio signal with the reconstructed drum part based on 128 neural networks

Для оценки эффективности применения исследуемых методов для поставленной задачи предлагается исследовать в совокупности несколько представлений результатов: спектрограммы восстановленных аудиосигналов, значения метрик MSE, MAE, и median absolute error (см. табл.), а также результат в аудиоформате. Анализируя полученные данные, можно прийти к выводу, что лучший результат получен при количестве нейронных сетей равном 32. Значения только метрики в нашем случае не информативно для оценки результата предсказания, так

как неясно, появились ли в аудиосигнале барабаны или возник кадровый шум. Именно поэтому, чтобы сделать вывод об эффективности используемых методов, следует принимать все представления результатов в совокупности.

Таким образом, как видно на рис. 6–9, при увеличении количества нейронных сетей результат работы приближается к ожидаемому, но при этом увеличивается уровень шума аудиосигнала.

Метрики для валидации результата
Metrics for result validation

	16	32	64	128
mean squared error	0,017456	0,003772	0.,009480	0,020648
mean absolute error	0,032182	0,014861	0,023788	0,036868
median absolute error	0,003299	0,003031	0,003105	0,004277

Заключение

Полученные в ходе работы результаты позволяют утверждать, что при наличии сигнала с искажением или с удаленной частью существует возможность построить сложную регрессионную модель для решения задачи восстановления оригинального сигнала при помощи свёрточных нейронных сетей. В предыдущих разделах показано, что существует взаимосвязь между поврежденными и оставшимися данными, которую можно аппроксимировать для восстановления сигнала.

В результате была разработана архитектура составной нейронной сети, состоящей из нескольких более простых сверточных сетей, количество которых может варьироваться. На основании полученных результатов можно утверждать, что при увеличении количества нейронных сетей в рамках решения той же задачи эффективность восстановления сигнала также увеличивается. Более того, полученные результаты показали высокий потенциал использования сверточных нейронных сетей для решения такой сложной задачи, как восстановление аудиосигнала.

Список литературы

1. **Mokrý O., Rajmic P.** Audio inpainting: Revisited and reweighted //IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020. Vol. 28. Pp. 2906–2918.
2. **Heo H. S., So B. M., Yang I. H., Yoon S. H., Yu, H. J.** Automated recovery of damaged audio files using deep neural networks //Digital Investigation. 2019. Vol. 30. Pp. 117–126.
3. **Kwon B., Gong M., Huh J. and Lee S.** Identification and Restoration of LZ77 Compressed Data Using a Machine Learning Approach. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018, pp. 1787–1790. DOI: 10.23919/APSIPA.2018.8659755
4. **Mousavi A., Patel A. B., Baraniuk R. G.** A deep learning approach to structured signal recovery. 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, 2015, pp. 1336–1343. DOI: 10.1109/ALLERTON.2015.7447163
5. **Zhang Z., Rao B. D.** Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning. IEEE Journal of Selected Topics in Signal Processing. 2011. Vol. 5, no. 5. Pp. 912–926. DOI: 10.1109/JSTSP.2011.2159773

6. **Bellasi D., Maechler P., Burg A., Felber N., Kaeslin H., Studer C.** Live demonstration: Real-time audio restoration using sparse signal recovery. 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, 2013, pp. 659–659. DOI: 10.1109/ISCAS.2013.6571929.
7. **Adler A. et al.** Audio inpainting //IEEE Transactions on Audio, Speech, and Language Processing. 2011. Vol. 20, no. 3. Pp. 922–932.
8. **Doshi S.** Various Optimization Algorithms For Training Neural Network. Towards Data Science. 2019 Jan. 13.
9. **Diederik P. Kingma and Jimmy Lei Ba.** Adam : A method for stochastic optimization. 2014. arXiv:1412.6980v9

References

1. **Mokrý O., Rajmic P.** Audio inpainting: Revisited and reweighted //IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020. Vol. 28. Pp. 2906–2918.
2. **Heo H. S., So B. M., Yang I. H., Yoon S. H., Yu H. J.** Automated recovery of damaged audio files using deep neural networks // Digital Investigation. 2019. Vol. 30. Pp. 117–126.
3. **Kwon B., Gong M., Huh J., Lee S.** Identification and Restoration of LZ77 Compressed Data Using a Machine Learning Approach. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018, pp. 1787–1790. DOI: 10.23919/APSIPA.2018.8659755
4. **Mousavi A., Patel A. B., Baraniuk R. G.** A deep learning approach to structured signal recovery. 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, 2015, pp. 1336–1343. DOI: 10.1109/ALLERTON.2015.7447163.
5. **Zhang Z., Rao B. D.** Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning. IEEE Journal of Selected Topics in Signal Processing, 2011. Vol. 5, no. 5, pp. 912–926. DOI: 10.1109/JSTSP.2011.2159773.
6. **Bellasi D., Maechler P., Burg A., Felber N., Kaeslin H., Studer C.** Live demonstration: Real-time audio restoration using sparse signal recovery. 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, 2013, pp. 659–659. DOI: 10.1109/ISCAS.2013.6571929
7. **Adler A. et al.** Audio inpainting //IEEE Transactions on Audio, Speech, and Language Processing, 2011. Vol. 20, no. 3, pp. 922-932.
8. **Doshi S.** Various Optimization Algorithms For Training Neural Network. Towards Data Science. 2019 Jan.13.
9. **Diederik P. Kingma and Jimmy Lei Ba.** Adam: A method for stochastic optimization. 2014. arXiv:1412.6980v9

Информация об авторах

Дементьева Кристина Игоревна, ассистент кафедры прикладной математики и кибернетики, аспирант, Сибирский государственный университет телекоммуникаций и информатики (Новосибирск, Россия)

Антон Андреевич Ракитский, кандидат техническиъ наук, доцент кафедры прикладной математики и кибернетики, Сибирский государственный университет телекоммуникаций и информатики (Новосибирск, Россия); научный сотрудник, Новосибирский государственный университет (Новосибирск, Россия); старший научный сотрудник, Институт вычислительной техники СО РАН (Новосибирск, Россия)

Information about the Authors

Kristina I. Dementyeva, Assistant of the Department of Applied Mathematics and Cybernetics, Postgraduate student, Siberian State University of Telecommunications and Informatics (Novosibirsk, Russian Federation)

Anton A. Rakitskiy, Candidate of Technical Sciences, Associate Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Informatics (Novosibirsk, Russian Federation); Researcher, Novosibirsk State University (Novosibirsk, Russian Federation); Senior Researcher, Institute of Computer Engineering SB RAS (Novosibirsk, Russian Federation)

Статья поступила в редакцию 10.06.2022;
одобрена после рецензирования 07.10.2022; принята к публикации 07.10.2022
The article was submitted 10.06.2022;
approved after reviewing 07.10.2022; accepted for publication 07.10.2022