Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.383

Духовник Андрей Николаевич

# Отделение голоса и звука из звуковых записей с использованием глубоких нейронных сетей

**АВТОРЕФЕРАТ**
на соискание академической степени
магистра информатики и вычислительной техники

по специальности 1-40 81 04 – Обработка больших объемов информации

Научный руководитель
Хотеев А.Л.
кандидат технических наук, доцент

Минск 2019

# КРАТКОЕ ВВЕДЕНИЕ

В данной работе рассмотрена задача выделения звука (акапелла) из звуковых аудиозаписей с использованием методов нейронных сетей, по аналогии с их применением к изображениям. Примером таких аудиозаписей могут служить live-записи на концертах, аудиозаписи сделанные в караоке или подобных приложениях.

Актуальность данной задачи подчеркивается возрастающим запросом к такого рода данным. Создаются запросы и темы на популярных ресурсах, например Reddit, где пользователи ищут доступ к исходных голосовым записям или к инструментам, которые позволяют выделять акапеллу. Основными недостатками существующих подходов является то, что они подразумевают либо изначальную раздельную запись голоса и музыки, либо способность незначительного уменьшения громкости инструментального компонента. С другой стороны, задача поставленная этой работе осложняется именно природой звуковых данных.

Существующих аналогов решения данной задачи с использованием нейронных сетей нет, так как основные методы, которые встречаются в популярных программах, например, Soundmane и пакетах/библиотеках, например, Librosa, основаны на усилении спектра в диапазоне от 400 до 2000 Hz, либо наоборот в заглушении остальной части спектра. Таким образом данная задача является актуальной на данный момент.

Основная сложность обработки звуковых аудиозаписей является то, что если изображения выглядят одинаково и на мониторе и в жизни, то со звуком все сложнее. Однако существуют подходы, которые позволяют нивелировать данные недостатки и использовать сходные методы для решения задач компьютерного зрения, для решения задач для аудио формата.

Долгое время для решения сходных задач использовались различные фильтры и другие классические методы, которые не давали достаточного эффекта. Однако с активным ростом уровня развития, и как следствие, популярности машинного обучения, а конкретно сверточных нейронных сетей, появилось огромное количество исследователей, применяющих подобные методы для любых задач. Таким образом попытка использовать данные методы для решения данных задач является натуральной.

# ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

## Цель и задачи исследования

*Целью* диссертационной работы является изучение методов применения нейронных сетей, методов глубокого обучения, для решения задачи выделения голоса из аудиозаписей. Разработка алгоритма для выделения голоса (акапеллы) из звуковых аудиозаписей, содержащих музыку, голос и шум, реализация алгоритма на выбранном языке программирования. Для достижения поставленной цели необходимо решить следующие задачи:

1 Ознакомиться с цифровыми методами представления и обработки звуковых данных.

2 Определить сферы применения методов машинного обучения для задач связанными с обработкой аудиозаписей, а именно методов, использующие нейронные сети, в том числе глубокого обучения. Провести аналогию между использованием методов нейронных сетей для решения задач связанных с изображениями, и задачами по обработки аудиофайлов.

3 Разработать многошаговый алгоритм для обработки звукозаписей и выделения голоса из них.

4 Реализовать алгоритм, реализовать нейронную сеть, собрать необходимые данные, обучить модель.

5 Провести экспериментальные исследования разработанного алгоритма и модели нейронной сети.

*Объектом* исследования являются методы обработки звуковых данных.

*Предметом* исследования является применение нейронных сетей для обработки звуковых данных, а именно выделения голоса из аудиозаписей.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования глубоких нейронных сетей для обработки звуковых записей по аналогии с применением нейронных сетей для обработки графических изображений. Существует возможность представлять звуковые данные в виде, аналогичному цифровому виду представления изображений, с помощью FFT. Также дополнением является схожесть задачи выделения голоса из аудиозаписей с задачей шумоподавления в аудиофайлах. Для достижения поставленной задачи необходимы данные, которые находятся в открытом доступе в сети интернет.

**Личный вклад соискателя**

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя А.Л. Хотеев, заключается в формулировке целей и задач исследования.

**Апробация результатов диссертации**

Основные положения диссертационной работы докладывались и обсуждались на международной научно-технической конференции молодых ученых «Новые материалы, оборудование и технологии в промышленности» (Могилев, Беларусь, 2018).

**Опубликованность результатов диссертации**

По теме диссертации опубликовано 2 печатные работ, из них 2 работы в сборниках трудов и материалов конференций.

**Структура и объем диссертации**

Диссертация состоит из введения, общей характеристики работы, пяти глав, заключения, списка использованных источников, списка публикаций автора, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, рассмотрены методы работы со звуком в цифровом представлении. Вторая глава посвящена нейронным сетям. Основным положениям, краткой истории развития, рассмотрении основных архитектур и описание сверточных нейронных сетей. В третьей главе рассмотрены основные используемые технологии. В четвертой главе описано исследовании поставленной задачи, обсуждены основные положения, предложен механизм сбора данных, а также представлен предлагаемый алгоритм выделения голоса из аудиозаписей. В пятой главе рассмотрены результат работы предложенного алгоритмы, сделаны выводы.

Общий объем работы составляет 61 страница, из которых основного текста 44 страницы, 21 рисунок, использованных источников состоящих из 21 наименование на 2 страницах и 2 приложения на 8 страницах.

# ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы, кратко затронута тема аналогов, и попытка решить данную задачу аналогично схожей в домене изображений.

В **первой главе** проведен анализ способам представления звука в цифровых системах. Затронута физическая сторона вопроса, а именно, что такое звук, звуковая волна, его распространение. Далее поднимается вопрос о цифровом представлении звука. Объяснена суть модуляции, что это влечет к неизбежной потере информации и артефактам, с которыми нужно бороться.

Остальная часть главы посвящена быстрому преобразованию Фурье: какая у этого метода математическая суть, как оно применяется для анализа звука и других сигналов. Затронуто логичное распространение преобразование Фурье на временной домен, что в свою очередь перерастает в спектрограммы или сонограммы. Также обращается внимание на необходимость использования различных оконных преобразований, например, окно Хэмминга, для устранения краевых эффектов.

**Вторая глава** посвящена нейронным сетям. Затрагивается история нейронных сетей. Подробно разбирается структура данных моделей машинного обучения, начиная с понятия нейрона, связи между нейронами, функции активации, заканчивая различными архитектурами, такими как: полностью связные нейронные сети, сверточные нейронные сети, рекуррентные и т.п.

Подробно описываются сверточные нейронные сети, как основной метод решения поставленной задачи. Представляется концепция сверточного слоя, ядер сверточного слоя, затронуты понятия пуллинг, паддинг и т.п.

В данного главе описывается схема обучения нейронной сети и приводятся проблемы, которые могут появляться в процессе обучения. Также описывается метод обратного распространения ошибки, как метод обучения. Заостряется внимание на методах градиентного спуска, как методах оптимизации.

В **третьей главе** рассматриваются используемые технологии. Вся работа была выполнена на языке программирования Python, который является de-facto стандартом для индустрии для машинного обучения и анализа данных. В качестве фреймворков для машинного обучения и глубокого

обучения использовались Tensorflow и Keras, которые являются наиболее популярными инструментами среди разработчиков и исследователей. Обучение нейронной сети занимает ощутимое время, поэтому была использована технология CUDA для распараллеливания обучения сети на GPU. Для анализа звуковых файлов и работы с ними была использована библиотека языка Python – Librosa, которая обладает богатым функционалом для считывания и визуализации звуковых файлов, имеет множество встроенных и уже реализованных фильтров, позволяет строить различные преобразования, получать амплитудно-спектральные характеристики, и строить спектрограммы, которые являются необходимым для работы шагом.

В **четвертой главе** представлен ход исследований и полученный алгоритм обработки аудиозаписей. Был приведен общие положения исследований, какие были выдвинуты гипотезы, которые привели к подобному решению. Представлен общий алгоритм обработки звука, состоящий из предобработки, модели нейронной сети, постобработки и восстановления звука. Предобработка включает в себя применение фильтров, процесс разделения аудио на части (разделенные по времени), вычисление спектрограммы по каждой из частей. Далее представлена нейронная сеть, которая является полностью сверточной, имеющий residual связи. Нейронная сеть запускается на каждой части звука, преобразованной в спектрограмму, а на ее выходе находится спектрограмма такой же последовательности, но уже без частей, которые отвечают за музыку и шум. После сети стоит еще один слой фильтров, который запускается на уже восстановленном аудиофайле.

Также был приведен процесс сбора данных и получения датасета, необходимого для тренировки сети. Красота такого датасета состоит в том, что можно как угодно перемешивать голоса и инструментальные части, увеличивая размер датасета, а также возможность решать задачу как self-supervised.

**Пятая глава** посвящена выводам, приведению результатов работы представленного алгоритма и нерешенных проблем. Были проанализированы результаты, таким образом сделан вывод, что данная модель может удалять музыку и шумы из аудиозаписей, тем самым оставляя только голос. Однако на аудиозаписи могут появляться некоторые артефакты, такие как металлический скрежет или эхо. Данные артефакты могут быть удалены как и применением дополнительных моделей, так и изменением текущего алгоритма, добавлением новых фильтром и изменением архитектуры сети, либо ее переобучением.

# ЗАКЛЮЧЕНИЕ

## Основные научные результаты диссертации

1 Задача выделения голоса из аудиозаписей, содержащих музыку и шум, была сформулирована как само-контролируемая (self-supervised) задача глубокого обучения с использованием нейронных сетей.

2 Предложен алгоритм обработки аудиозаписей, состоящий из препроцессинга и постпроцессинга, механизма разделения аудио файла на части, которые обрабатываются независимо, полностью сверточной нейронной сети, которая получая спектрограмму аудиосигнала, стирает на нам элементы, отвечающие за шум и музыку, и механизм восстановления аудиозаписей из спектрограммы.

3 Исследованы результаты на различных аудиозаписях, найдены примеры, которые хорошо обрабатываются с помощью алгоритма, и найдены проблемные зоны. Из основных проблем можно выделить электронный скрежет, который появляется на аудиофайлах, барабанные партии, которые алгоритм в текущем виде игнорирует, и появляющееся эхо, которого появляется после обработки. Данные проблемы можно решать различными способами, из них можно выделить применение специальных фильтров, изменение архитектуры сети или созданий специальной модели для убирания эхо и барабанов.

4 Показано, что данная задача может решаться как self-supervised задача машинного обучения с использованием нейронных сетей, и неплохие результаты можно достичь с небольшими вычислительными мощностями и относительно небольшим датасетом. Но для достижения результатов, в которых человеческое ухо не будет замечать никаких артефактов, необходимы большее количество данных и серьезные вычислительные мощности.

## Рекомендации по практическому использованию результатов

1 Полученные результаты формируют теоретическую и практическую базу для разработки ПО компьютерных систем для решения задачи выделения голоса из аудиозаписей. Данные программы могут работать в реальном времени, выполняться на ПК и мобильных устройствах.

2 Была подтверждена гипотеза о том, что данную задачу можно решать по аналогии с задачей шумоподавления изображений, но с использованием спектрограмм аудиозаписей. Данные результаты могут быть использованы для более масштабных исследований, с большими вычислительными ресурсами и количеством данных.

3 Результаты работы могут использоваться для создания различных приложений. Например, задача выделения голоса из аудиозаписей может быть полезна для компаний, которые производят приложения Караоке, а также для внедрения в различные файлообменники или стриминговые сервисы для выделения голоса из видеозаписей, например, из живой записи концерта.

# СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Духовник А.Н., Бутома В.С.: Отделение голоса от музыки из аудиозаписей с использованием методов глубокого обучения / Сборник материалов международной научно-технической конференции молодых ученых "Новые материалы, оборудование и технологии в промышленности". – Могилев, 2018. – 151 с.

2-А. Духовник А.Н., Бутома В.С.: Метод улучшения результатов сверточных нейронных сетей методом зеркального нахлеста изображения / Сборник материалов международной научно-технической конференции молодых ученых "Новые материалы, оборудование и технологии в промышленности". – Могилев, 2018. – 152 с.