УДК 004.8 **Информационные технологии**

Сайгин Андрей Александрович, *студент 2 курса магистратуры,*
Мордовский государственный университет имени Н. П. Огарёва,
Россия, Саранск
e-mail: andrexsai@mail.ru
Плотникова Наталья Павловна, *кандидат технических наук, доцент,*
Мордовский государственный университет имени Н. П. Огарёва,
Россия, Саранск

ПОДГОТОВКА ДАННЫХ ДЛЯ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ, РЕШАЮЩИХ ЗАДАЧУ РАЗДЕЛЕНИЯ ИСТОЧНИКОВ СИГНАЛА

Аннотация: Звук играет большую роль в жизни человека. Распознавание звуков – это один базовых инстинктов, позволявших людям избегать опасности. С помощью звука оценивается обстановка вокруг человека, находящаяся за пределами его поля зрения. Поэтому качество звука играет большую роль. Только из качественного звука, без артефактов и лишних шумов можно получить необходимую информацию. Но зачастую в звуковых записях присутствуют посторонние звуки, которые мешают распознаванию информации, из чего возникла задача фильтрации посторонних шумов. Формально, данная проблема называется разделение источников звука. Она заключается в восстановлении или реконструкции одного или нескольких исходных сигналов, которые в результате линейного или сверточного процесса смешаны с другими сигналами. Для упрощения фильтрации и шумоподавления звука в наше время начали применять алгоритмы машинного обучения.

Ключевые слова: Звук, фильтрация, разделение источников звука, нейронная сеть, модель, данные, частота дискретизации, WAV, сжатие, датасет, аугментация, нормализация, амплитудная характеристика, частотная характеристика, преобразование Фурье, спектрограмма.

Abstract: Sound plays a big role in a person's life. Sound recognition is one of the basic instincts that allowed people to avoid danger. With the help of sound, the situation around a person outside his field of vision is assessed. Therefore, sound quality plays a big role. Only from high-quality sound, without artifacts and unnecessary noise, you can get the necessary information. But often there are extraneous sounds in sound recordings that interfere with the recognition of information, which led to the task of filtering extraneous noise. Formally, this problem is called audio source separation. It consists in restoring or reconstructing one or more original signals that are mixed with other signals as a result of a linear or convolutional process. To simplify filtering and noise reduction of sound, machine learning algorithms have begun to be used in our time.

Keywords: Sound, filtering, audio source separation, neural network, model, data, sampling rate, WAV, compression, dataset, augmentation, normalization, amplitude response, frequency response, Fourier transform, spectrogram.

Введение

В последние годы машинное обучение обрело большую популярность. Объясняется это тем, что алгоритмы машинного обучения способны решать задачи, которые решить стандартным путем практически невозможно. Самым важным аспектом, благодаря которому обучение в принципе возможно, являются данные, на которых алгоритм обучается. Машинное обучение сильно зависит от данных.

Проблема в том, что во всех массивах данных есть изъяны. Модель, построенная на некачественных данных, будет неточной, а результаты её работы не получится использовать. Поэтому до обучения модели нужно провести подготовку исходных данных: очистить и привести к нужному формату.

Формат данных

Составление датасета для обучения моделей, решающих задачу разделения источников сигнала, происходит следующим образом. Собирается набор чистых

сигналов, которые выступают в роли эталона. Это могут быть студийные записи речи, дорожки фильмов, музыка. Также, собирается набор «шумов». Это могут быть записи с улиц, помещений, сэмплы из банков звуков. Стоит учитывать, что типы шумов должны быть разнообразными и сбалансированными, иначе нейронная сеть обучится фильтровать только один вид «шума». Набранные «шумы» аугментируются с эталонными записями – выполняется наложение одних звуковых дорожек на другие. Таким образом получается два набора данных – чистый звук и соответствующий зашумленный.

Изначальные записи должны быть стандартизированы. В ходе данного процесса выравнивается амплитуда сигнала посредством центрирования среднего значения амплитуды. Процедура стандартизации амплитуды необходима по той причине, что исходные данные в звуковых записях не нормированы (могут быть очень малы или очень велики). По этой причине задавать универсальные пороговые значения невозможно. Стандартизация решает эту проблему [1]. После данной процедуры можно выполнять наложение «шумов».

При обучении может возникнуть проблема сходимости нейронной сети, что отрицательно скажется на результате обучения. Для решения этой проблемы элементы обучающей выборки должны иметь одинаковую длину, и она должна быть не большой. Поэтому ранее полученные звуковые дорожки разделяются на отрезки длиной в 2 секунды. В ходе разделения могут возникнуть пустые дорожки, которые не содержат полезную информацию. Такие фрагменты удаляем из набора.

Звуковая информация имеет важную характеристику, которая влияет на работу нейронной сети. Это частота дискретизации – частота взятия отсчётов непрерывного по времени сигнала при его дискретизации (в частности, аналого-цифровым преобразователем) [2]. При разной частоте дискретизации будут браться разные отсчеты непрерывного сигнала, из-за чего один и тот же сигнал, дискретизированный с разной частотой будет записан по-разному. Если набирать данные из сигналов с разными частотами дискретизации, то датасет

получится неоднородным, из-за чего нейросеть не сможет правильно обучиться. Поэтому приведем все звуковые дорожки к одинаковой частоте дискретизации. Самым распространенным значением частоты дискретизации в настоящее время является 48 кГц.

В настоящее время звук чаще всего записывается методом стереофонии. Это метод записи и воспроизведения звука, при котором создаётся иллюзия «звуковой перспективы» с сохранением направлений на разные источники звука. Это достигается за счёт использования бинаурального эффекта и одновременной передачи звуковой информации по двум и более независимым каналам, в отличие от монофонической звукопередачи, когда звук передаётся по единственному каналу [3]. Получается, что звуковая запись имеет несколько каналов, каждый из которых несет свою информацию. Существует несколько спецификаций стереозвука, в которых количество каналов может достигать 8. Обрабатывать такую информацию становится намного сложнее, потому что для каждого канала потребуется отдельный вход нейронной сети. Кроме того, при обучении может возникнуть ситуация, что на каком-либо из каналов вообще не будет шумов, из-за чего дальнейшее ее применение будет не эффективным. Поэтому лучше всего составить датасет из одноканальных монозаписей. При практическом использовании обученной модели для многоканальных записей можно будет применять ее для каждого канала отдельно.

Для хранения звуковой информации в компьютере было разработано большое количество цифровых аудиоформатов. Для данного понятия существуют различные определения. Они связаны с форматом представления и форматом файла [4]. Формат представления зависит от способа квантования, его разрядности и частоты дискретизации. Формат файла определяет структуру представления данных и используемый аудиокодек, при помощи которого производится сжатие аудиоданных. Выделяют три группы звуковых форматов:

- без сжатия;
- сжатие без потерь – закодированные данные могут быть однозначно восстановлены;

- сжатие с потерями – раскодированные данные несущественно отличаются от исходных.

Поскольку каждый кодек по-разному кодирует аудиосигналы, то и их представление так же будет отличаться от формата к формату, что так же приводит к неоднородности данных. Поэтому нужно выбрать единый формат аудиозаписей, который будет обрабатываться фильтром. Он должен быть форматом без сжатия, потому что при конвертировании из одного кодека в другой и обратно повышается риск потери нужных данных. Кроме того, формат должен поддерживать все ранее описанные ранее параметры, а также быть кроссплатформенным. Под все описанные признаки подходит формат файлов WAV. Это медиаконтейнер, разработанный совместно компаниями Microsoft и IBM, для хранения несжатого звука. Формат файла является открытым, поэтому он может быть открыт любым плеером. Он поддерживает множество битовых разрешений, частот дискретизации и каналов звука. Недостатком формата является большой размер файла.

В итоге получаем, что для обучения будут использованы аудиозаписи, хранящиеся в формате WAV, длительностью 2 секунды, с разрешением 16 бит, частотой дискретизации 48000 Гц, одноканальные.

Источники датасетов

Поскольку задача разделения сигналов и интеллектуального удаления шумов не нова, то уже был создан ряд датасетов, на основе которых можно проводить обучение моделей для решения данной задачи.

Чаще всего для при обучении используется датасет, собранный Microsoft для соревнования Deep Noise Suppression (DNS) Challenge [5]. По сути, это сборник из большого числа датасетов с чистым звуком, сэмплов с шумами и программный код для их слияния. Соревнование уже проходит пятый раз, поэтому каждый год датасет обновляется. Чистые записи представляют из себя речь на разных языках (английском, немецком, французском, итальянском испанском, русском) и с разной эмоциональной окраской. Общий объем чистых записей составляет 827 Гб, а шумов – 58 Гб.

Для публикации звуковых датасетов существует сайт OpenSLR. OpenSLR – это сайт, посвященный размещению речевых и языковых ресурсов, таких как учебные корпуса для распознавания речи и программное обеспечение, связанное с распознаванием речи [6]. Ресурс задумывался как удобное место для размещения созданных наборов данных в открытом доступе. На сайте на данный момент выложено 133 различных датасетов с речью на разных языках.

Набор данных WSJ0 Hipster Ambient Mixtures (WHAM!) берет за основу набор данных wsj0-2mix с уникальной фоновой сценой шума [7]. Звук шума был собран в различных городских районах в районе залива Сан-Франциско в конце 2018 года. Окружающая среда в основном состоит из ресторанов, кафе, баров и парков. Звук был записан с помощью бинаурального микрофона Apogee Sennheiser на штативе на высоте от 1,0 до 1,5 метров от земли.

Большая база звуковых данных собрана на сайте AudioSet, созданный командой Google Research [8]. Данные набирались и размечались на основе видео с сайта YouTube. Было собрано около 5800 часов аудиозаписей, распределенных по 527 классам. В датасете содержатся музыка, человеческая речь, голоса животных, звуки с улиц и помещений.

Кроме готовых датасетов можно самостоятельно набрать данные из открытых источников. Чистые сигналы должны отличаться студийным качеством записи. Раньше такие аудиозаписи требовали дорогого оборудования, поэтому доступ к ним был не свободным и на них лежали правовые ограничения. Сейчас звукозаписывающее оборудование, дающее высокое качество результата, стало доступнее как в плане цены, так и в плане установки и настройки, поэтому возросло количество звукового контента, выкладываемого в открытых источниках.

Самым крупным открытым источником информации является Internet Archive. Internet Archive – некоммерческая организация, основанная в 1996 году в Сан-Франциско американским программистом Брюстером Кейлом. Главной заявленной целью Архива является предоставление всеобщего доступа к накопленной в Интернете информации. Коллекция АИ состоит из множества

подколлекций архивированных веб-сайтов, оцифрованных книг, аудио- и видеофайлов, игр, программного обеспечения [9]. В Архиве сохранено терабайты разнородной информации. По состоянию на 2022 год на серверах Архива сохранено около 15 миллионов аудиозаписей. К ним относятся аудиокниги, музыка, записи радиоэфиров, подкасты и многое другое.

В роли шумов можно использовать различные сэмплы, применяемые при видеомонтаже. Источников бесплатных звуковых эффектов в Интернете много, правда надо учитывать, что не все они распространяются под свободными лицензиями.

Подготовка данных

Подготовка звуковых файлов состоит из этапов предобработки, аугментации и получения характеристик записей.

Предобработка звука включает в себя следующие шаги:

- загрузка файла;
- разделение каналов;
- разделение потоков на блоки;
- нормализация блока;
- сохранение с необходимыми параметрами.

Операцию предобработки необходимо выполнить и с чистыми записями, и с шумами. По итогу получатся два набора файлов.

Аугментация заключается в том, что необходимо взять чистую запись и наложить на случайную запись с шумом, после чего результат нормализуется и сохраняется с необходимыми параметрами. Имена чистых записей и соответствующих зашумленных должны совпадать.

Поскольку звуковая запись представляет из себя сигнал, то его можно визуализировать в виде графиков и получить его характеристики. Самыми важными для оценки являются амплитудная характеристика, частотная характеристика и спектрограмма.

Амплитудная характеристика – это представление сигнала во временной области. Она показывает амплитуду звуковой волны, изменяющуюся во

времени. Пример амплитудной характеристики изображен на рис. 1.

Рис. 1. Построение амплитудных характеристик сигналов

Амплитудная характеристика не очень информативна, так как говорят только о громкости аудиозаписи. Чтобы лучше понять звуковой сигнал, необходимо преобразовать его в частотную область. Представление сигнала в частотной области сообщает нам, какие разные частоты присутствуют в сигнале. Для ее построения используется преобразование Фурье. Пример частотной характеристики изображен на рис. 2.

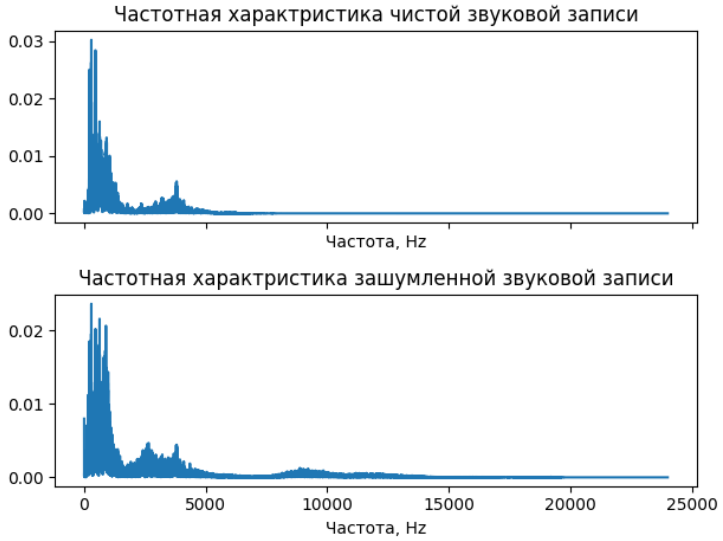

Рис. 2. Построение частотных характеристик сигналов

Но зачастую требуется знать, в какие моменты времени проявляли себя те или иные частоты. Поэтому требуется наличие характеристик как по времени, так и по частотам. Для решения данной задачи существуют спектрограммы – представление частот сигнала во времени. На графике представления спектрограммы одна ось представляет время, вторая ось представляет частоты, а цвета представляют величину (амплитуду) наблюдаемой частоты в конкретный момент времени. Пример спектрограммы изображен на рис. 3.

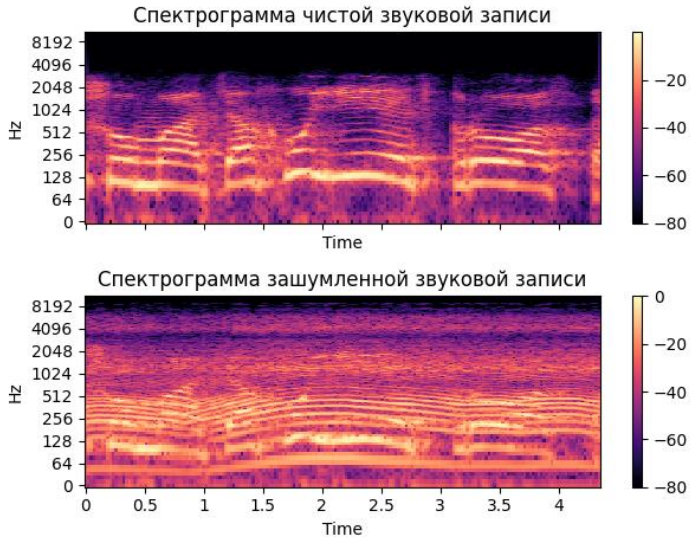

Рис. 3. Построение спектрограмм сигналов

По графикам можно провести сравнение чистого и зашумленного сигналов, определить места проявления шумов, самые выделяющиеся частоты.

Заключение

Распознавание звуков – это один базовых инстинктов, позволявших людям избегать опасности. И сейчас звуки продолжают играть большую роль в нашей жизни. С помощью звука оценивается обстановка вокруг человека, находящаяся за пределами его поля зрения.

Качество звука играет большую роль. Только из качественного звука, без артефактов и лишних шумов можно получить необходимую информацию.

Нейронные сети являются эффективным инструментом для фильтрации звуковой информации от посторонних шумов. Уже существует несколько решение на базе различных архитектур – от классических рекуррентных сетей, до U-net с применением механизма self-attention. Но эффективность подобных инструментов в первую очередь зависит от данных, на которых алгоритмы обучались.

Библиографический список:

1. Лялин С. Г. Метод шумоподавления в речевых сигналах с помощью нейронной сети / С. Г. Лялин // Advanced Science – 2019. – № 2 – С. 32–38.
2. Гольденберг Л. М. Цифровая обработка сигналов: Учебное пособие для вузов / Л. М. Гольденберг – 2-е изд. – М.: Радио и связь, 1990. – 256 с.
3. Высоцкий М. З. Системы кино и стереозвук / М. З. Высоцкий. – М.: «Искусство», 1972. – 336 с.
4. Батов С. Форматы звуковых файлов часть 1 / С. Батов // «Звукорежиссер» — информационно-технический журнал, предназначенный для специалистов в области профессиональных звуковых технологий. – 1999. – № 8.
5. Cutler R. ICASSP 2021 Deep Noise Suppression Challenge / C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, S. Srinivasan – Текст: электронный // arXiv.org – URL: https://arxiv.org/abs/2009.06122 – Дата публикации: 26 октября 2020.
6. OpenSLR: сайт. URL: http://www.openslr.org/index.html (дата обращения: 28.12.2022).
7. Wichern G. WHAM!: Extending Speech Separation to Noisy Environments / G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, J. Le Roux – Текст : электронный // arXiv.org – URL : https://arxiv.org/abs/1907.01160 – Дата публикации: 2 июля 2019.
8. AudioSet: сайт. URL: https://research.google.com/audioset/ (дата обращения: 28.12.2022).

9. Internet Archive: сайт – 1996. URL: https://archive.org/ (дата обращения: 28.12.2022).