TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.01 «Информатика и вычислительная техника»
Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Алгоритм обнаружения речевой активности в акустическом сигнале с применением сверточных нейронных сетей

УДК 004.932.1.032.26

Студент

Группа	ФИО	Подпись	Дата
8ВМ83	Тепляков Андрей Борисович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОИТ ИШИТР	Спицын Владимир Григорьевич	д. т. н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН ШБИП	Конотопский В. Ю.	к. э. н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Горбенко М. В.	к. т. н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
09.04.01 "Информатика и вычислительная техника", профиль "Компьютерный анализ и интерпретация данных"	Спицын Владимир Григорьевич	д. т. н.		

Томск – 2020 г.

Запланированные результаты обучения по основной образовательной программе подготовки магистров 09.04.01 «Информатика и вычислительная техника», ИШИТР НИ ТПУ, профиль «Компьютерный анализ и интерпретация данных»

Код	Результат обучения (выпускник должен быть готов)
Общепрофессиональные компетенции	
Р1	Самостоятельно приобретать, и применять математические, естественнонаучные, социально-экономические и профессиональные знания в области современных информационно-коммуникационных технологий для решения междисциплинарных инженерных задач.
Р2	Разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач.
Р3	Применять на практике новые научные принципы и методы исследований. Демонстрировать способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями.
Р4	Разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем.
Р5	Анализировать и оценивать уровни своих компетенций в сочетании со способностью и готовностью к дальнейшему образованию и профессиональной мобильности. Активно владеть одним из иностранных языков на уровне социального и профессионального общения, применять специальную лексику и профессиональную терминологию языка.

Р6	Осуществлять эффективное управление разработкой программных средств и проектов. Эффективно работать, как член и руководитель группы, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре.
Профессиональные компетенции	
Р7	Разрабатывать стратегии проектирования, критерии эффективности и ограничения применимости сверточных нейронных сетей и методов вычислительного интеллекта для разработки программно-алгоритмических систем анализа больших объемов данных.
Р8	Планировать и проводить теоретические исследования и компьютерные эксперименты в области создания программных систем интеллектуального анализа больших объемов данных.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.01 «Информатика и вычислительная техника»
Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

_____ _____ _____
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

Студенту:

Группа	**ФИО**
8ВМ83	Теплякову Андрею Борисовичу

Тема работы:

Алгоритм обнаружения речевой активности в акустическом сигнале с применением сверточных нейронных сетей	
Утверждена приказом директора	№140-46/с от 19.05.2020 г.

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе:	Наборы аудиозаписей, находящиеся в открытом доступе и содержащие как речевую активность, так и звуковые сигналы, которые можно отнести к категории шум.
Перечень подлежащих исследованию, проектированию и разработке вопросов:	– аналитический обзор предметной области; – изучение существующих подходов к обнаружению речевой активности;

	– проектирование и реализация собственного алгоритма – оценка полученных результатов и практической применимости разработанного алгоритма; – рассмотрение вопросов финансового менеджмента, ресурсоэффективности и ресурсосбережения в соответствующем разделе; – рассмотрение вопросов социальной ответственности в соответствующем разделе.
Перечень графического материала:	– UML диаграмма классов модуля формирования обучающей, валидационной и тестовой выборок – UML диаграмма классов модуля обучения сверточной нейронной сети
Консультанты по разделам выпускной квалификационной работы	
Раздел	**Консультант**
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Конотопский В. Ю.
Социальная ответственность	Горбенко М. В.
Названия разделов, которые должны быть написаны на русском и иностранном языках:	
Модуль формирования выборок	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОИТ ИШИТР	Спицын Владимир Григорьевич	Д. Т. Н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ83	Тепляков Андрей Борисович		

Реферат

Выпускная квалификационная работа содержит 95 страниц, 22 рисунка, 17 таблиц, 37 литературных источников, 1 приложение.

Ключевые слова: автоматическая обработка речи, обнаружение речевой активности, машинное обучение, глубокое обучение, сверточные нейронные сети.

Объектом исследования являются методы глубокого обучения, применяемые в задачах автоматической обработки речи. Предметом исследования являются алгоритмы обнаружения речевой активности с применением сверточных нейронных сетей.

Целью работы является разработка и реализация алгоритма обнаружения речевой активности во входном акустическом сигнале для отделения человеческой речи от фонового шума или тишины.

В исследовании представлены обзоры предметной области, существующих аналогов, а также инструментов, с помощью которых возможна разработка данного алгоритма. Спроектированы и реализованы модули для формирования выборок и обучения сверточной нейронной сети. Проведено сравнение реализованного алгоритма с аналогами по точности и скорости обнаружения речевой активности.

К области применения алгоритма относятся технологии голосового пользовательского интерфейса.

Запланированным развитием результатов исследования является оптимизация предложенного алгоритма для увеличения его производительности без потери качества за счет использования возможностей графических ускорителей.

Содержание

Введение

В настоящее время голосовой пользовательский интерфейс приобрел широкую популярностью. Такие голосовые помощники как Cortana, Siri, Google Assistant, Алиса ежедневно обрабатывают значительное количество запросов, повышая комфорт пользователя при выполнении рутинных операций [1]. Значительное количество как зарубежных, так и российских банков используют технологию интерактивного голосового ответа, снижая расходы на персонал.

В автомобильной промышленности приложения громкой связи и голосового управления позволяют водителю взаимодействовать с людьми и самой машиной во время вождения, не отвлекаясь от дорожного движения.

Также усовершенствованная обработка речевого сигнала может помочь людям с нарушениями слуха. Современные слуховые аппараты усиливают желаемый речевой сигнал и подавляют мешающие шумовые компоненты [2].

Хотя существуют различные варианты применения технологий обработки речевого сигнала, разработанные алгоритмы сталкиваются с общей проблемой: необходимо обнаружить присутствие речи в акустическом сигнале, который зачастую искажен шумом.

Целью данной работы является разработка и реализация алгоритма обнаружения речевой активности (Voice Activity Detection) во входном акустическом сигнале для отделения человеческой речи от фонового шума или тишины.

Данную цель можно разделить на следующие задачи:

1. изучение предметной области;
2. поиск существующих аналогов алгоритма, рассмотрение их преимуществ и недостатков;
3. обзор инструментов, с помощью которых возможна разработка данного алгоритма;

4. реализация алгоритма обнаружения голосовой активности;
5. сравнение реализованного алгоритма с аналогами по точности детектирования речевой активности.

1. Теоретические сведения

1.1. Представление звукового сигнала в вычислительной технике

Как известно, звук - это физическое явление, которое представляет собой продольное распространение механических колебаний в различных средах. При этом звукозаписывающее устройство, часто называемое микрофоном, преобразует звуковые волны в изменения напряжения. Если микрофон подключен к звуковой карте, то напряжение можно измерять через равные промежутки времени (с заданной частотой дискретизации) и каждое значение преобразовывать в двоичное число. Данный процесс называется квантованием по уровню звука и выполняется аналого-цифровым преобразователем на звуковой карте, после чего серия двоичных чисел может быть сохранена в виде звукового файла. На рисунке 1 представлен пример квантования по уровню и времени непрерывного сигнала.

Под частотой дискретизации понимают количество измерений разности потенциалов аналоговой звуковой волны, взятых в секунду. Эта частота измеряется в Герцах (Гц).

Звуковая карта может воссоздать сохраненный звук с помощью цифро-аналогового преобразователя, то есть последовательность двоичных чисел преобразуется обратно в изменяющееся напряжение, которое вызывает вибрацию динамика для воспроизведения звука [3].

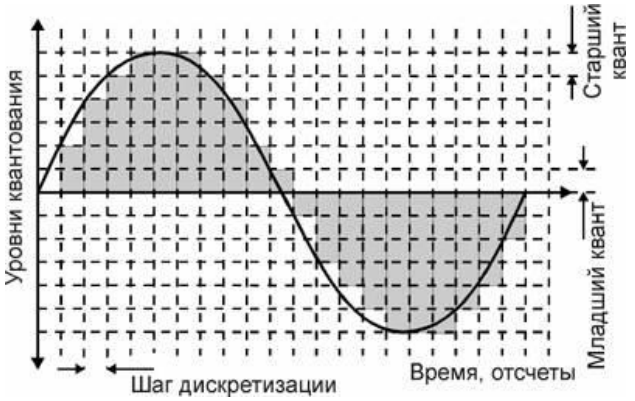


Рисунок 1 – Квантование по уровню и времени непрерывного сигнала

1.2. Признаки, позволяющие отличить речевую активность от шума

На рисунке 2 изображен звуковой сигнал, на котором участки, где график пунктирной линии принимает значение 1, соответствуют речевой активности, определенной слушателем. В остальное время записана различная шумовая активность (хлопки, щелчки, шуршание бумагой).

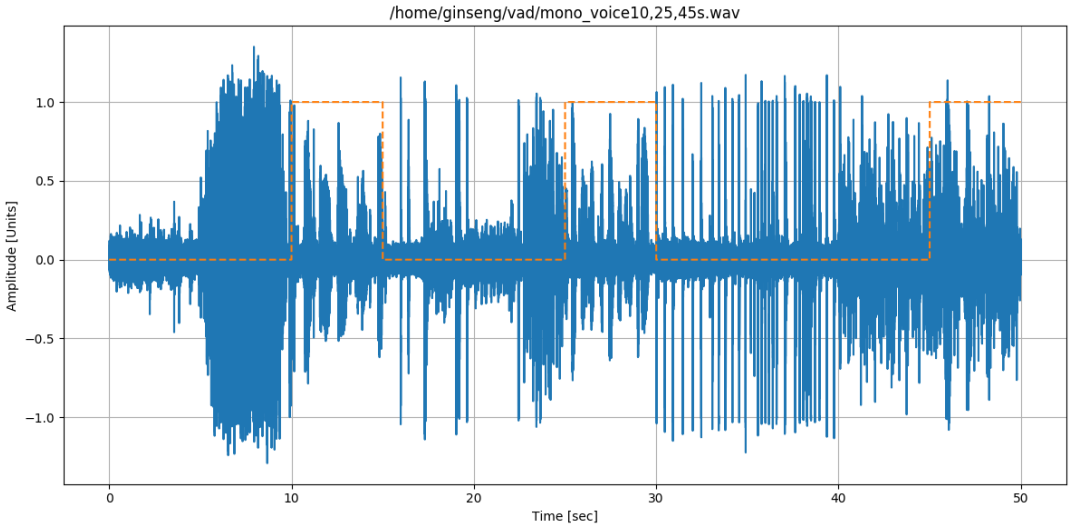


Рисунок 2 – Звуковой сигнал с речевой и шумовой активностями

В данном случае, в отличии, например, от задачи классификации изображений, человеку достаточно сложно выделить признаки, по которым можно понять, что является речью, а что - нет. Если же наложить на исходный сигнал белый шум, как показано на рисунке 3, то при прослушивании речь все еще различима, однако визуально выделить паттерны речевой активности практически невозможно.

А ведь такой сценарий с низким отношением сигнал-шум встречается чаще всего, и детектор речевой активности должен справляться с данной ситуацией.

Для изучения акустического сигнала принято рассматривать его порции методом скользящего окна, как показано на рисунке 4.

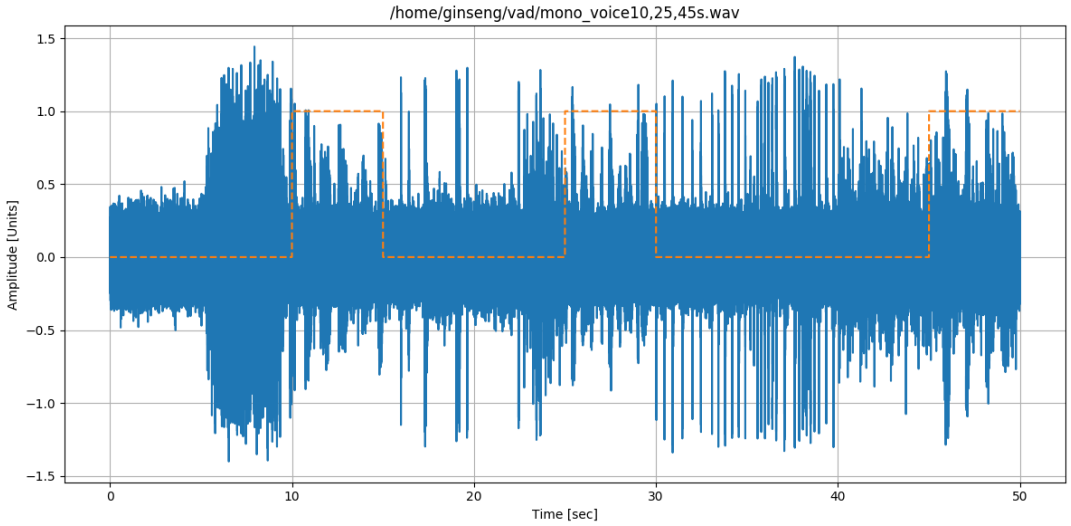


Рисунок 3 – Звуковой сигнал с добавлением белого шума

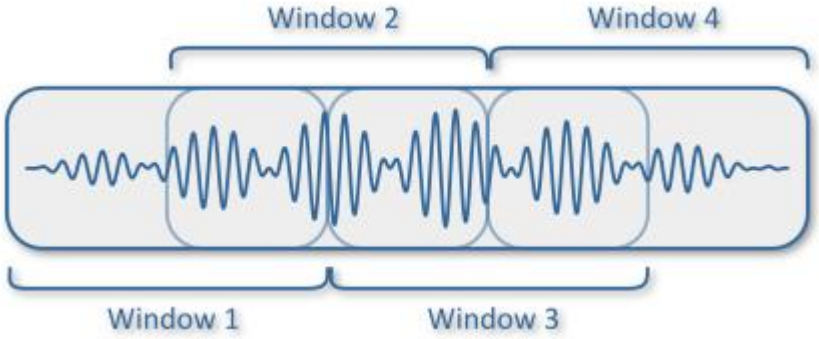


Рисунок 4 – Применение скользящего окна к сигналу

Как можно заметить на рисунке выше, окна идут с наложением друг на друга, при этом чем больше наложение, тем выше временное разрешение. В то же время увеличивается время выполнения алгоритма. Общепринятой практикой является рассмотрение окна длительностью около 100 мс. В среднем столько человек произносит один слог.

Первой характеристикой, часто используемой для обнаружения речевой активности во временной области сигнала, является краткосрочная энергия (Short Term Energy), вычисляемая по следующей формуле:

$$STE(i) = \frac{1}{N}\sum_{n=1}^{N} {x_n}^2(i),$$

где $x_n(i), n = 1, 2, \dots, N$ – последовательность из N отсчетов сигнала для i-го окна.

В данном случае для обнаружения речи выдвигается предположение, что ее компоненты демонстрируют более высокие значения мощности по сравнению с фоновым шумом. Поэтому если значение краткосрочной энергии выше некоторого порога, то участок сигнала относится к речевой активности. Во многих сценариях допущение об увеличении мощности оправдано из-за эффекта Ломбарда [4], согласно которому говорящий повышает голос в шумной обстановке. Однако фиксированное пороговое значение требует априорных знаний об уровнях шума и речи. Нормализация мощности увеличивает разделимость между компонентами речи и шума, однако нестационарные помехи, такие как ударные шумы, вызывают ложные срабатывания детектора речевой активности на основе мощности.

Бóльшая устойчивость к шумам достигается при отображении сигнала из временной области в частотную при помощи дискретного преобразования Фурье (ДПФ, Discrete Fourier Transform).

Данное преобразование является фундаментальным в цифровой обработке сигналов с приложениями для частотного анализа, быстрой свертки, обработки изображений. Кроме того, существуют алгоритмы эффективного вычисления ДПФ, называемые быстрыми преобразованиями Фурье, первый из которых предложили Кули и Тьюки [5].

В общем случае, прямое преобразование осуществляется по следующей формуле:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} = \sum_{n=0}^{N-1} x_n \cdot [\cos(2\pi kn/N) - i \cdot \sin(2\pi kn/N)], \qquad (k = 0, \dots, N-1).$$

Здесь N – количество отсчетов дискретного сигнала; x_n – измеренные значения сигнала; X_k – комплексные амплитуды синусоидальных сигналов, слагающих исходный сигнал; являются выходными данными для прямого

преобразования и входными для обратного; поскольку амплитуды комплексные, то по ним можно вычислить одновременно и амплитуду, и фазу; k – индекс частоты [6].

Отсчеты для преобразования выбираются тем же методом скользящего окна, описанным выше, а затем отображаются на графике, который называется спектрограммой. Это визуальное представление спектра частот сигнала, который изменяется со временем, при этом интенсивность пикселя показывает мощность амплитуды определенной частотной составляющей сигнала в определенный момент времени. Пример спектрограммы звукового файла представлен на рисунке 5.

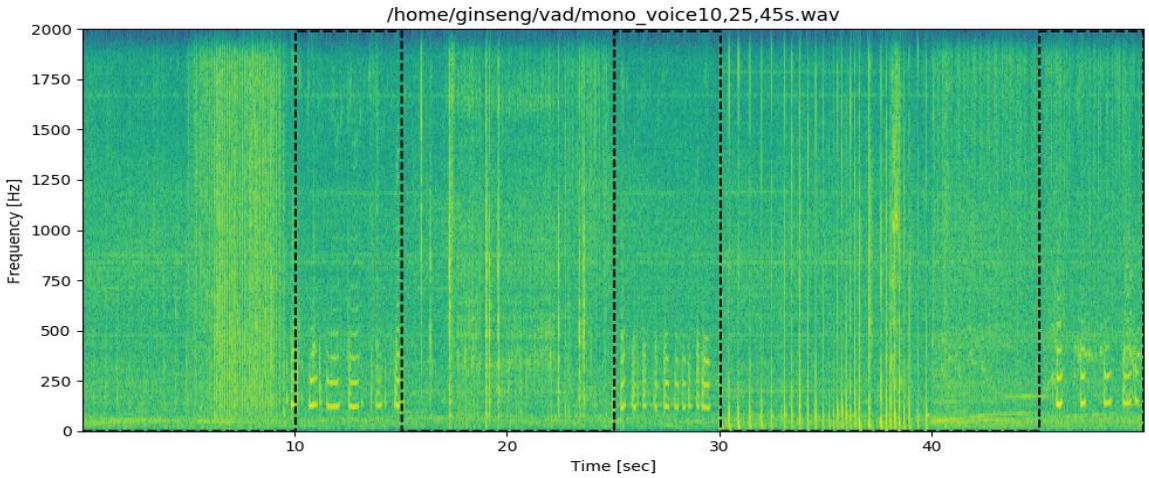


Рисунок 5 – Спектрограмма исходного звукового сигнала

Основная частота, создаваемая речевым трактом человека, варьируется от 80 Гц до 180 Гц для мужчин и от 160 Гц до 260 Гц для женщин [7]. Это утверждение подтверждается при анализе рисунка 5. В низкочастотной области сигнала можно заметить речевые паттерны. Самое полезное свойство представления сигнала в частотной области заключается в том, что отличительные характеристики речевой активности сохраняются и при наложении белого шума, что можно увидеть на рисунке 6.

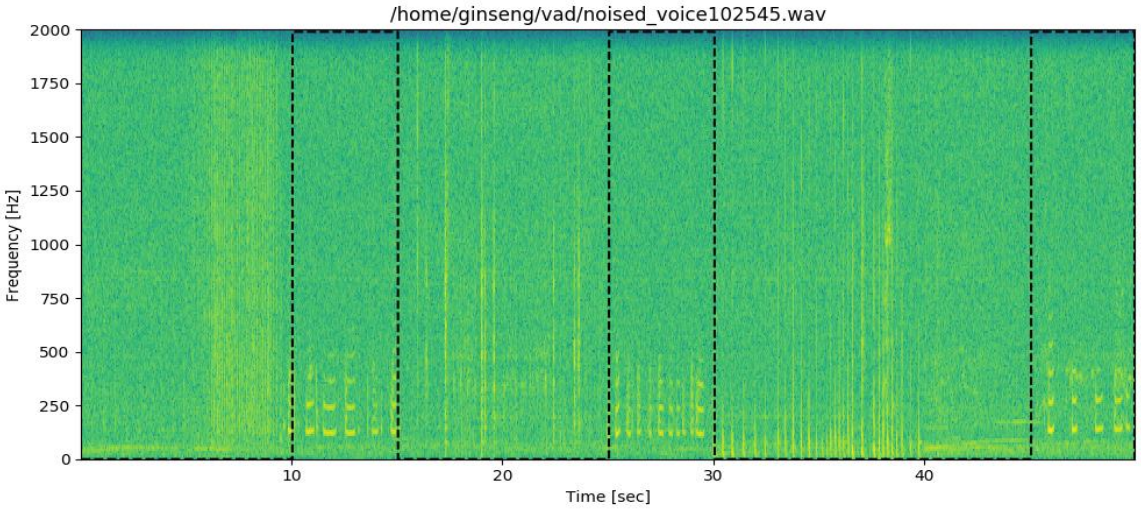


Рисунок 6 – Спектрограмма звукового сигнала с добавлением белого шума

Зная данные особенности речи, эксперты выделяют несколько частотных признаков.

Спектральная плоскостность (Spectral Flatness Measure) является мерой шумности спектра и её можно использовать для того, чтобы различить речевую и шумовую активности. Значение данного признака рассчитывается по следующей формуле:

$$SFM(i) = \frac{(\prod_{n=1}^{N} X_n(i))^{\frac{1}{N}}}{\frac{1}{N}\sum_{n=1}^{N} X_n(i)},$$

где $X_n(i), n = 1, 2, \dots, N$ – спектр сигнала из N частот после преобразования Фурье для i-го окна.

При спектральной плоскостности близкой к нулю можно утверждать, что сигнал имеет несколько мощных гармоник, и если они сосредоточены в области низких частот, то можно предположить, что это голос. В обратном случае, при близости данного коэффициента к единице, можно предположить, что данный сигнал близок к белому шуму (спектр частот распределен равномерно) [8].

Другим признаком речевой активности является коэффициент полосы частот спектра (Spectrum Frequency Band Ratio), который представляет собой

отношение суммы магнитуд определенной полосы частот к сумме всех магнитуд спектра [9]. Он вычисляется по формуле:

$$SFBR(i) = \frac{\sum_{n=T}^{K} X_n(i)}{\sum_{n=1}^{N} X_n(i)},$$

где T и K – границы диапазона частот, которые в случае присутствия в сигнале речевой активности должны иметь наибольшую мощность. При исследованиях в данной работе были установлены 80 и 1000 Гц соответственно.

Другие характеристики для обнаружения речевой активности можно изучить в [10].

Исследованные характеристики изображены на рисунке 7 для звукового файла, представленного ранее, при этом на нижнем графике представлено поточечное произведение трех признаков, описанных выше.

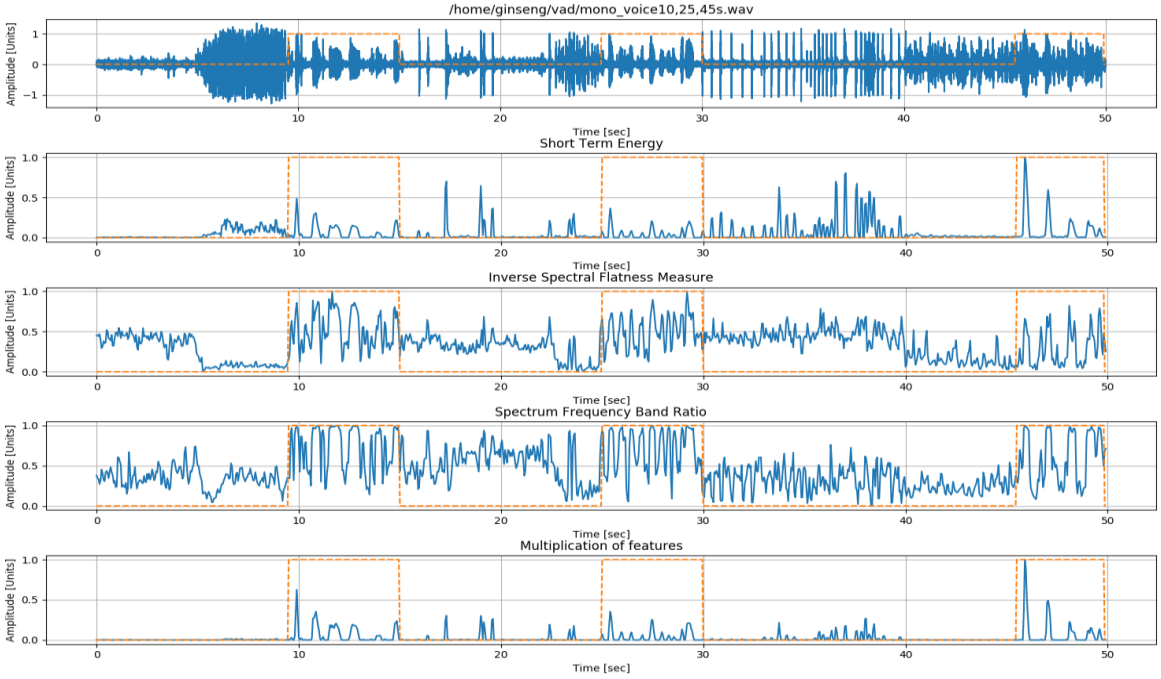


Рисунок 7 – Признаки речевой активности для исходного звукового сигнала

В результате анализа представленных выше графиков можно заключить, что между значениями признаков и участками с речевой активностью есть зависимость, однако она имеет сложный нелинейный характер. В связи с этим для обнаружения речевой активности следует

воспользоваться моделями машинного обучения. В данной работе алгоритм обнаружения речевой активности строится на основе искусственной нейронной сети.

Также следует отметить, что каждый из описанных выше признаков – это единственное значение для скользящего окна. Поэтому их основным недостатком является ложно положительное срабатывание на шумы, схожие по спектральным характеристикам с речью. В этой работе входными данными для модели являются сами спектрограммы, так как они содержат более полную информацию об особенностях речевой активности. Благодаря этому модель в ходе обучения должна получить способность к выделению признаков для обнаружения речевой активности.

1.3. Искусственные нейронные сети

Искусственная нейронная сеть (ИНС) - это математическая модель, описывающая систему соединенных между собой искусственных нейронов, а также реализации этой модели. ИНС построены в некотором смысле по образу и подобию биологических нейронных сетей.

В общем случае, нейронной сетью решается задача аппроксимации функции $f*$ таким образом, что вектор входных значений x отображается в вектор целевых значений y: $y=f(x)$. Под обучением понимают процедуру подбора таких параметров W ИНС, которые дают наиболее эффективную аппроксимацию функции $f*$: $y = f(x, W)$.

В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными и выходными данными, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть возвращает корректный в некотором смысле результат на основании неполных или частично искаженных данных.

Структурной единицей ИНС является нейрон. С точки зрения математики, он представляет собой взвешенную сумму входных сигналов, которая является аргументом нелинейной функции, называемой активационной функцией или функцией активации (рисунок 8).

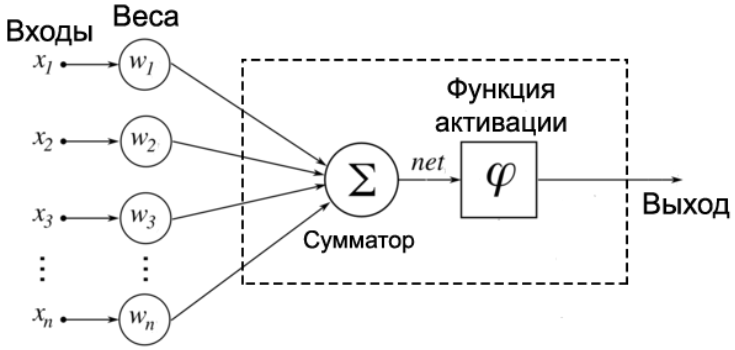


Рисунок 8 – Искусственный нейрон

В ИНС нейроны объединяются в слои в зависимости от особенностей их функционирования. Слои принято нумеровать слева направо. Первый слой называется входным. Нейроны этого слоя принимают поступающие входные данные и передают их следующему слою. Последний слой называется выходным, так как он формирует результаты работы нейронной сети. Слои, расположенные между входным и выходным слоем, называются скрытыми или внутренними. В этих слоях происходит основная обработка данных. В случае, если выходные сигналы нейронов одного слоя передаются всем нейронам следующего, такой слой называется полносвязным (Fully Connected Layer).

При обучении ИНС для определения корректности полученного результата работы используется функция потерь. В задаче обучения с учителем функция потерь определяет различие между исходными значениями и предсказаниями модели.

В силу того, что в данной работе акустический сигнал отображается в частотную область и представляется в виде спектрограммы, следует использовать разновидность ИНС, которая называется свёрточной нейронной сетью (СНС). В задачах обработки изображений она позволяет добиться

значительного улучшения точности и снижения вычислительной сложности по сравнению с полносвязными сетями.

Данный тип моделей глубокого обучения используется для обработки данных с известной топологией, например, изображений, которые можно рассматривать как двумерную сетку пикселей. Слово “сверточная” в названии указывает на то, что в модели используется математическая операция свертки вместо умножения матриц, по крайней мере на одном из слоёв, который называется свёрточным [11].

Идея свёрточных нейронных сетей была изложена в конце прошлого века Яном Лекуном [12], однако значительный скачок в развитии данных моделей произошел в 2012 году, когда Алекс Крижевски выиграл ежегодное соревнование Large Scale Visual Recognition Challenge по классификации изображений на наборе данных ImageNet [13].

Работу данной модели можно описать следующим образом. Каждый нейрон слоя получает входной сигнал от локального рецептивного поля в предыдущем слое. Под рецептивным полем понимается слой нейронов, воспринимающих внешние сигналы и передающих их следующему слою.

Иными словами, при прямом проходе входных данных через модель получается, что каждый нейрон выполняет операцию свертки (конволюции) некоторой области предыдущего слоя, которая определяется множеством нейронов, связанных с данным нейроном.

Пример применения операции свертки с ядром 3x3 и шагом 2 к изображению с размером 7x7 показан на рисунке 9.

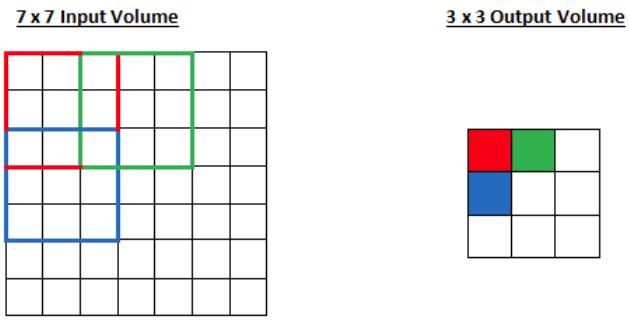


Рисунок 9 – Пример применения операции свертки

Помимо сверточных слоев в топологии модели могут быть слои субдискретизации и полносвязные слои. Слои субдискретизации выполняют функции уменьшения размерности пространства карт признаков. Данный слой позволяет ускорить дальнейшие вычисления. Субдискретизации возможна благодаря тому, что для данной модели важно не столько значение признака, сколько сам факт его наличия.

Пример операции подвыборки максимальных значений (max pool) с ядром 2x2 и шагом 2 показан на рисунке 10.

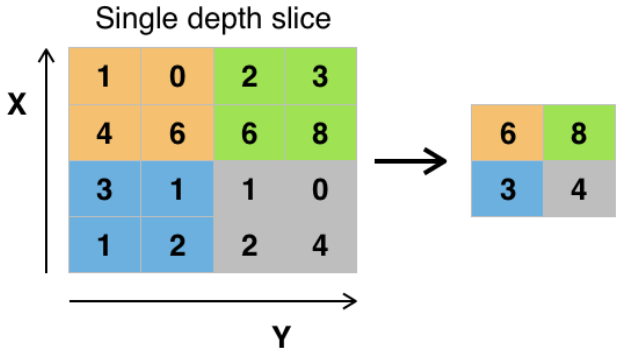


Рисунок 10 – Пример операции подвыборки

Выходной слой сверточной нейронной сети, как правило, всегда полносвязный. Все три вида слоев могут чередоваться в произвольном порядке. Это позволяет составлять карты признаков из карт признаков, что на практике даёт возможность перейти от рассмотрения конкретных особенностей входных данных к распознаванию абстрактных сущностей.

При обучении нейронных сетей обычно используется метод обратного распространения ошибки (Backpropagation) или его модификации.

2. Обзор существующих методов обнаружения речевой активности

2.1. Речевой кодек G.729 Annex B

Исследование данного вопроса следует начать с алгоритма, используемого в узкополосном речевом кодеке G.729, применяемом в целях эффективного цифрового представления речи, которая передается по телефонной связи в сети Интернет (VoIP). Использование алгоритма обнаружения голосовой активности обусловлено тем, что речь содержит множество пауз, которые не следует передавать по каналу связи и которые можно сгенерировать для слушателя.

В качестве расширения речевого кодека G.729 международный комитет ITU-T разработал G.729 Annex B с целью поддержки прерывистой передачи посредством обнаружения активности речи, анализа фонового шума и генерации комфортного шума. Кодек G.729B принимает сигнал длительностью 10 мс, после чего решение о наличии или отсутствии голосовой активности получается при рассмотрении четырех параметров: разность энергий всего диапазона частот; разность энергий диапазона низких частот; искажение спектра; разность частоты переходов через ноль. Блок обновления параметров шума основан на схеме авторегрессии первого порядка. Они обновляются, если разница энергии всего диапазона меньше заданного фиксированного порога [14].

Каждый, кто пользовался телефонной связью, представляет недостаток данного алгоритма именно со стороны обнаружения речевой активности, то есть эффективно удаляются только паузы в речи.

2.2. WebRTC VAD

Другим алгоритмом обнаружения речевой активности является WebRTC VAD. В целом, WebRTC – это проект с открытым исходным кодом, поддерживаемый Google, Mozilla, Opera и другими организациями. Цель

проекта предоставление браузерам и мобильным приложениям функций связи в реальном времени с помощью простых API.

В основе алгоритма лежит модель машинного обучения без учителя под названием гауссова смесь. На вход модели подаются участки аудиозаписи 10, 20 или 30 мс с частотой дискретизации 8000, 16000, 32000 или 48000 Гц, которые затем отображаются в частотную область преобразованием Фурье. Каждый участок в свою очередь разбивается на 6 частотных диапазонов: 80 - 250 Гц, 250 - 500 Гц, 500 - 1000 Гц, 1000 - 2000 Гц, 2000 - 3000 Гц, 3000 - 4000 Гц. Далее происходит вычисление энергии для каждого диапазона и общей энергии участка. Если общая энергия выше некоторого порогового значения, то для каждого диапазона применяется гауссова смесь в предположении, что есть два класса - шум и голос. Затем находится отношение правдоподобия того, что данный диапазон - голос, после чего принимается решение по пороговому значению, называемому агрессивность. Оно может быть равно 0 (алгоритм склонен чаще относить шум к речевой активности), 1, 2 и 3 (алгоритм склонен чаще относить речевую активность к шуму).

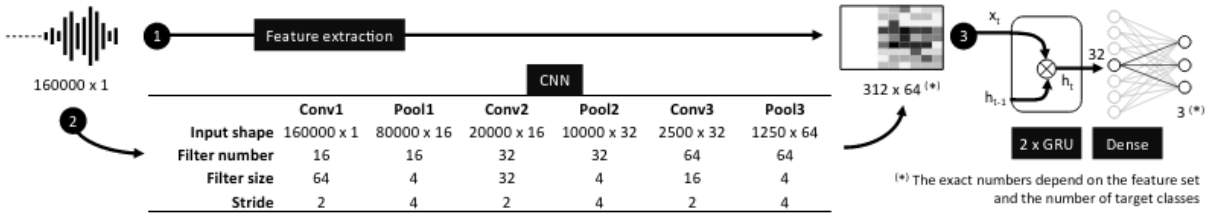
В репозитории [15] представлен проект, позволяющий использовать данный алгоритм с помощью языка программирования Python. При проверке работоспособности алгоритма были получены неудовлетворительные результаты для сигналов с низким отношением сигнал-шум, результаты приведены в разделе 5.

2.3. VadNet

К алгоритмам, использующим методы глубокого обучения относится VadNet [16]. Авторы утверждают, что в прошлом качество работы алгоритмов машинного обучения сильно зависело от представления данных. Поэтому хорошо продуманные признаки играли ключевую роль в задачах распознавания речи и паралингвистики. В связи с этим инженеры проделали большую работу по исследованию сложных акустических признаков. Однако

с появлением глубоких нейронных сетей стало возможным автоматически выводить более высокие абстракции из простых спектральных представлений или даже учиться непосредственно из необработанных сигналов.

Архитектура разработанной нейронной сети схематично представлена на рисунке 11.



	Conv1	Pool1	Conv2	Pool2	Conv3	Pool3
Input shape	160000 x 1	80000 x 16	20000 x 16	10000 x 32	2500 x 32	1250 x 64
Filter number	16	16	32	32	64	64
Filter size	64	4	32	4	16	4
Stride	2	4	2	4	2	4

Рисунок 11– Архитектура VadNet

Модель принимает в качестве входных данных необработанный аудиосигнал, который передается через три сверточных слоя. Полученные карты признаков обрабатываются двухслойной рекуррентной сетью. Последний слой является полносвязным, к выходу которого применяется функция Softmax. Так входной сигнал отображается в вероятности его отнесения к голосу и шуму.

К недостаткам данного подхода следует отнести вычислительную сложность из-за значительного размера входа и использования рекуррентной нейронной сети.

В связи с тем, что все найденные аналоги не имеют желаемых характеристик, следует разработать свою модель обнаружения речевой активности.

3. Обзор инструментов реализации алгоритма

Основным инструментом является язык программирования Python из-за удобства быстрой прототипизации благодаря большому количеству фреймворков как для работы с аудиофайлами (в частности для получения спектрограмм), так и для глубокого обучения. Одной из наиболее удобных и функциональных интегрированных сред разработки для Python является PyCharm, которая и использовалась в данной работе. Далее проведен сравнительный анализ фреймворков глубокого обучения [17].

3.1. TensorFlow

Данный фреймворк с открытым исходным кодом разработан компанией Google на языках программирования Python и C++. TensorFlow заслужил популярность как в академической, так и в индустриальной сфере, ведь такие крупные компании как Uber, DeepMind, Dropbox и AirBnB выбрали этот фреймворк для глубокого обучения моделей.

Эксперты выделяют несколько преимуществ TensorFlow:

- существует большое количество руководств для начинающих и качественная документация;
- поддерживается крупным сообществом независимых программистов и техническими компаниями;
- поддерживает распределенное обучение на нескольких машинах;
- TensorFlow Lite обеспечивает вывод с низкой задержкой для мобильных устройств.

Также необходимо сказать о недостатках данного продукта компании Google. Во-первых, в сравнении с фреймворками CNTK и MXNet, TensorFlow несколько уступает по скорости работы в эталонных тестах. Во-вторых, он имеет более высокий входной порог для начинающих, чем PyTorch.

3.2. MXNet

Фреймворк для глубокого обучения разработан Apache и поддерживает несколько языков программирования, в том числе Python, Julia, C++, R и JavaScript. Проекты с использованием MXNet реализуют Microsoft, Intel и Amazon Web Services. Среди преимуществ можно выделить:

- обеспечивает продвинутую поддержку GPU;
- имеет высокопроизводительное императивное API;
- обеспечивает легкую поддержку моделей;
- обладает высокой масштабируемостью;
- обеспечивает поддержку множества языков программирования.

Главным недостатком MXNet является отсутствие крупного сообщества разработчиков, которое есть у TensorFlow и PyTorch.

3.3. PyTorch

Данный фреймворк является преемником Python для библиотеки Torch, написанной на Lua. Он был разработан Facebook и используется в разработках Twitter, Salesforce, Оксфордского Университета и многими другими компаниями и университетами.

PyTorch имеет несколько сильных сторон:

- более понятный процесс создания модели, чем на TensorFlow;
- поддержка популярных инструментов для отладки программы, таких как pdb или PyCharm Debugger;
- наличие значительного количества моделей, в том числе, предварительно обученных, а также готовых модульных частей с возможностью их комбинирования;
- возможность распределенного обучения с версии 0.4.

Из-за отсутствия серьезных недостатков данный фреймворк выбран для реализации алгоритма обнаружения голосовой активности.

4. Модуль формирования выборок

4.1. Используемые наборы данных

Набор данных для обучения должен включать в себя как можно более разнообразную голосовую активность индивидов. Для этих целей были использованы открытые наборы данных Common Voice и Voxforge.

Common Voice - это набор звуковых файлов, которые содержат записи, сделанные посетителями веб-сайта, на которых зачитывается текст из ряда общедоступных источников, таких как сообщения блогов, отправленные пользователями, старые книги, фильмы и другие публичные данные. Набор данных используется для обучения и тестирования систем автоматического распознавания речи [18].

Проект VoxForge был создан для сбора звуковых файлов с речью и для использования в системах распознавания речи с открытым исходным кодом, таких как ISIP, HTK, Julius и Sphinx. Речевой корпус доступен в формате GNU General Public License [19].

Также для обучения модели необходимы примеры не речевой активности. Для этой цели используются наборы данных ESC-50 и Urban Noises [20].

Набор данных Environmental Sound Classification представляет собой маркированную коллекцию из 2000 аудиозаписей окружающей среды, пригодных для сравнительного анализа методов классификации звуков окружающей среды. Набор данных состоит из записей продолжительностью 5 секунд, организованных в 50 семантических классов (по 40 примеров на класс), условно разбитых на 5 основных категорий: животные; звуки природы и воды; человеческие неречевые звуки; бытовые домашние звуки; городские шумы [21].

4.2. Проектирование модуля формирования выборок

При проектировании программного обеспечения для формирования выборок следует выделить два основных интерфейса: для работы с наборами данных и для обработки аудиофайлов.

Создание первого позволяет унифицировать способ получения пути до аудиозаписи в файловой системе и скрыть специфическую для каждого набора данных структуру каталогов.

С помощью интерфейса обработчика возможна реализация специфических операций над аудиофайлом, которые будут рассмотрены далее. При этом следует обеспечить возможность для построения цепочки из обработчиков в целях аккумулирования результатов преобразований.

Основным обработчиком является тот, который реализует операции по формированию спектрограмм, так как в данной работе они рассматриваются в качестве входных данных для сверточных нейронных сетей. Как известно, аудиозаписи могут иметь различную продолжительность, поэтому получаемые из них спектрограммы разделяются на изображения одинакового размера методом скользящего окна. Детали процесса формирования спектрограмм рассмотрены в следующем подпункте.

В данный момент следует рассмотреть подход к разметке полученных изображений, которая обычно является одним из наиболее трудоемких этапов подготовки обучающей выборки. В данном случае необходимо каждой спектрограмме присвоить метку 1, если известно, что на данном участке присутствует речевая активность, в противном случае - метку 0. В работе используется следующий подход к разметке. Категория, к которой следует отнести спектрограмму, определяется набором данных, из которого был взят аудиофайл.

В случае разметки наборов данных с шумами не возникает никаких проблем, так как все спектрограммы, полученные из аудиосигнала, имеют метку 0. В свою очередь для речевой активности характерно наличие участков

тишины из-за пауз между словами, спектрограммы которой ошибочно получат метку 1. Данный недостаток призван устранить обработчик, который предварительно удаляет тишину из аудиозаписи.

Описанных на данном этапе обработчиков достаточно, чтобы составить обучающую выборку из спектрограмм с речевой и шумовой активностями, однако существенный недостаток будет заключаться в следующем. В наборах данных Common Voice и VoxForge спикеры зачитывают различные отрывки в непосредственной близости от микрофона, тогда как в большинстве реальных сценариев использования хотелось бы иметь алгоритм, робастный к шумовому воздействию.

С этой целью следует аугментировать аудиозаписи с удаленной тишиной различными шумами из соответствующих наборов данных. За реализацию необходимой функциональности отвечает соответствующий обработчик.

Таким образом, для наборов данных с речевой активностью следует выстроить следующую последовательность из обработчиков: первоначально произвести удаление тишины, затем аугментировать аудиосигнал, после чего сформировать спектрограммы. Для наборов данных с шумами следует сразу производить спектрограммы из аудиозаписи.

4.3. Реализация модуля формирования выборок

Архитектура программного обеспечения для формирования выборок представлена на UML диаграмме классов, которая отображена на рисунке 12.

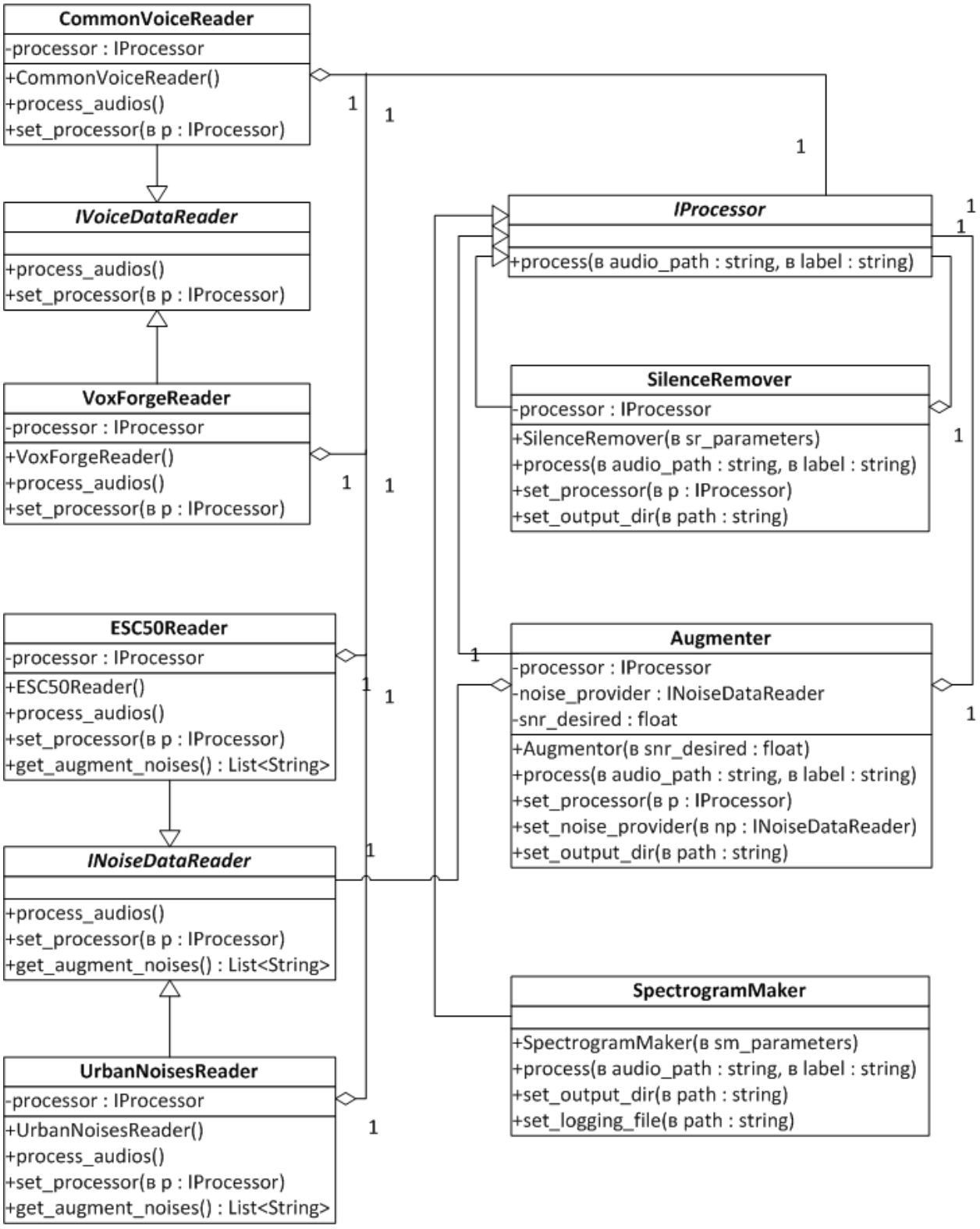


Рисунок 12 – UML диаграмма классов модуля

На диаграмме присутствуют абстрактные классы IVoiceDataReader и INoiseDataReader, которые предоставляют интерфейсы для чтения наборов данных с голосами и шумами соответственно. Их разделение обосновано тем,

что наборы данных с шумами должны передавать аудиофайлы для аугментации голосовой активности.

Классы CommonVoiceReader и VoxForgeReader реализуют интерфейс IVoiceDataReader, а именно методы для установки обработчика и преобразования каждого из аудиофайлов. При этом характер обработки, которая будет осуществляться в методе process_audios, определяется реализацией интерфейса IProcessor, а конкретно методом process.

В предыдущем подпункте были выделены три вида обработчиков, которые также можно увидеть на диаграмме. При этом для создания последовательно выполняющихся операций по обработке классы SilenceRemover и Augmentor позволяют установить дополнительные обработчики через метод set_processor.

Процесс по подготовке выборки начинается с того, что из аудиофайлов, содержащихся в наборах данных с речевой активностью, удаляется тишина, за что в данной архитектуре отвечает класс SilenceRemover. В нем реализовано вычисление признаков, описанных в разделе 1. Из него известно, что такие признаки речевой активности, как краткосрочная энергия STE и коэффициент полосы частот спектра SFBR позволяют достаточно успешно удалять тишину из аудиосигнала, который был записан при отсутствии посторонних шумов. При этом пороговые значения данных признаков подбираются для каждого оратора при прослушивании.

Затем, аудиозаписи с удаленной тишиной аугментируются, за что отвечает класс Augmentor. Под аугментацией в данной работе понимается сложение амплитуд сигналов с речевой и шумовой активностями для каждого отсчета. Для этого необходимо выполнение четырех операций.

Во-первых, должен быть установлен член данных класса Augmentor noise_provider, который является реализацией интерфейса INoiseDataReader и возвращает список аудиозаписей шумов для аугментации при вызове метода

get_augmentation_noises. При этом аудиофайлы случайно выбираются из каждой категории, имеющейся в наборе данных.

Во-вторых, для корректного сложения амплитуд необходимо, чтобы сигналы имели одинаковую частоту дискретизации. Изменению подвергается аудиозапись с шумами, частота дискретизации которой приводится к тому же значению, что и у аудиофайла с речевой активностью.

В-третьих, два складывающихся аудиосигнала чаще всего имеют разную продолжительность, тогда как для аугментации всего аудиофайла с речевой активностью они должны быть равны. Поэтому, в случае недостаточной продолжительности шума, он дублируется необходимое количество раз. Если же шум имеет большую продолжительность изначально или в результате дублирования, то избыточная часть сигнала не используется.

В-четвертых, амплитуды обоих сигналов чаще всего значительно отличаются. При их сложении возможны ситуации, когда на тихую речевую активность накладывается сильный шум. В таком случае, в действительности речь не будет слышна, однако ее разметка будет говорить об обратном. Также возможны случаи, в которых амплитуда шумового сигнала значительно меньше, чем у сигнала с речью, что приводит к слишком незначительным искажениям речевой активности. Чтобы избежать подобных ситуаций необходимо контролировать уровень шумового сигнала. Для этого вводится коэффициент k, на который умножаются амплитуды зашумляющего сигнала. Он вычисляется по формуле:

$$k = \sqrt{\frac{SNR}{SNR_{desired}}},$$

где $SNR_{desired}$ – желаемое значение отношения сигнал-шум в результате преобразования; SNR – исходное значение отношения сигнал-шум. Оно определяется следующим образом:

$$SNR = \frac{STE(x_V)}{STE(x_N)},$$

где x_V – амплитуды сигнала с речевой активностью; x_N – амплитуды зашумляющего сигнала; STE – краткосрочная энергия сигнала.

На рисунках 13 и 14 представлены примеры аугментации аудиосигнала исходно «слабыми» и «сильными» шумами соответственно. На каждом из рисунков в первом ряду приведен результат наложения исходного шума, а во втором – результат наложения шума, умноженного на коэффициент k.

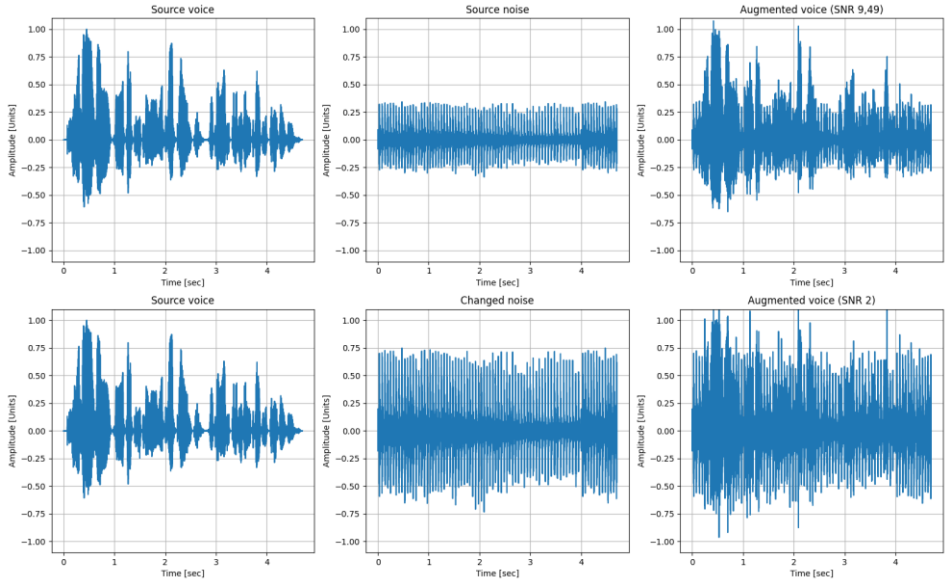


Рисунок 13 – Пример аугментации с исходно «слабым» шумом

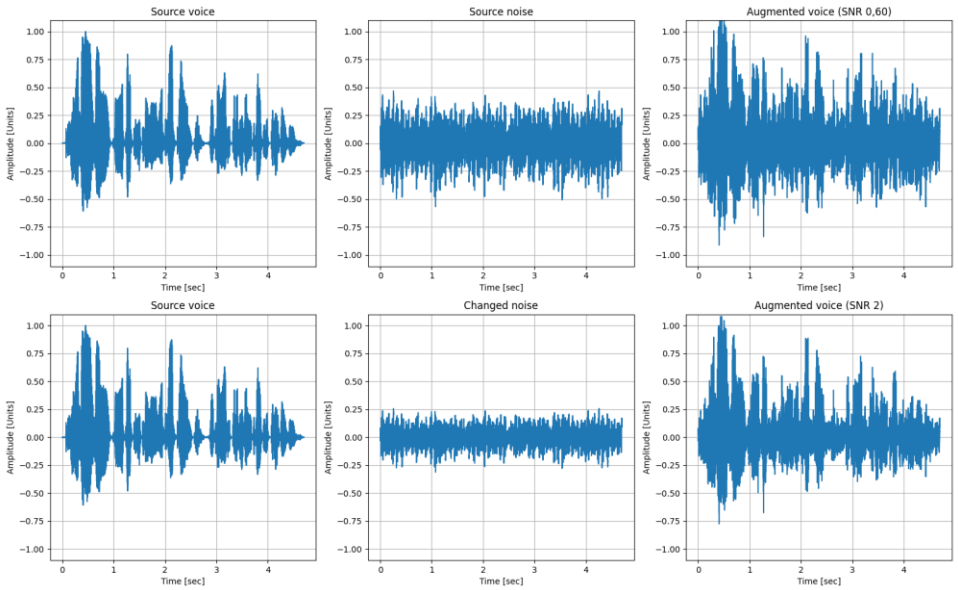


Рисунок 14 – Пример аугментации с исходно «сильным» шумом

Таким образом, в результате аугментаций удается увеличить выборку, а также добиться ее разнообразия.

Финальным обработчиком для аудиосигнала с речевой активностью и единственным для шумов является SpectrogramMaker, в котором реализуется следующее. Первоначально каждый звуковой файл проходит через процесс децимации на частоту дискретизации 4000 Гц. После этого из аудиосигнала формируется спектрограмма с параметрами: размер окна 400 отсчетов, интервал перекрытия 396 отсчетов. Таким образом, по оси абсцисс 1 пиксель соответствует 1 миллисекунде, а по оси ординат отображены частоты от 0 до 2000 Гц с шагом 10 Гц.

Входными изображениями для модели становятся изображения в оттенках серого 128 на 128 пикселей, то есть участки спектрограммы сигнала продолжительностью 128 миллисекунд и с мощностями гармоник сигнала на частотах от 0 до 1270 Гц с шагом 10 Гц. При этом для преобразования спектрограммы в изображение происходит нормировка, при которой минимальное значение амплитуды гармоники приравнивается 0, а максимальное - 255. Примеры спектрограмм можно увидеть на рисунке 15.

Рисунок 15 – Спектрограммы для обучения модели (слева - речь, посередине - дрель, справа - пианино)

4.4. Сценарий использования разработанного модуля

Для формирования выборок были использованы:

- 180 аудиозаписей из Common Voice;
- 175 аудиозаписей из VoxForge;

- 1600 аудиозаписей из ESC-50;
- 4000 аудиозаписей из Urban Noises.

Следует отметить, что из наборов данных с шумами были удалены категории, которые могут содержать речевую активность. В результате аугментации количество примеров с речевой активностью увеличивается в 8 раз за счет 7 категорий из набора данных Urban Noises и использования белого шума.

Все аудиозаписи разделяются для подготовки обучающей, валидационной и тестовой выборок в соотношении 70%, 10% и 20% соответственно. При этом аудиозаписи, относящиеся к тестовой выборке, аугментируются с разным значением желаемого SNR. В связи с этим в данной работе исследуемые модели и аналог тестируются на двух выборках: с низким воздействием помех (желаемый SNR равен 8) и с высоким уровнем шумовой активности (желаемый SNR равен 2). Для обучающей и валидационной выборок желаемый SNR также равен 2.

В результате работы модуля было сформировано 191191 спектрограмма для обучения, 27080 спектрограмм для валидации, и по 53213 спектрограмм в каждой из двух тестовых выборок. При этом выборка достаточно сбалансирована, так как 35% от общего количества спектрограмм относятся к категории речевой активности.

5. Реализация алгоритма обнаружения голосовой активности

5.1. Проектирование модуля для обучения модели

Целью работы данного модуля является обучение свёрточной нейронной сети, способной детектировать речевую активность на спектрограммах звукового сигнала. Для достижения поставленной цели необходимо следующее:

- подавать изображения на входной слой модели и выполнять прямой проход;
- передавать значения с выходного слоя и оригинальные метки в функцию потерь и выполнять обратный проход;
- используя метод градиентного спуска, оптимизировать параметры модели.

Данная последовательность действий выполняется до того момента, пока все обучающие примеры не будут показаны заданное количество раз. Качество работы обученной модели определяется по критериям, которые описаны в следующем подпункте.

Пользуясь функциональностью фреймворка PyTorch для решения поставленных задач спроектирована архитектура модуля для обучения модели, UML диаграмма классов которой представлена на рисунке 16.

Центральным в данном модуле является класс Runner, а именно его метод train, так как в нем реализована логика обучения. В качестве параметров данный метод принимает следующее:

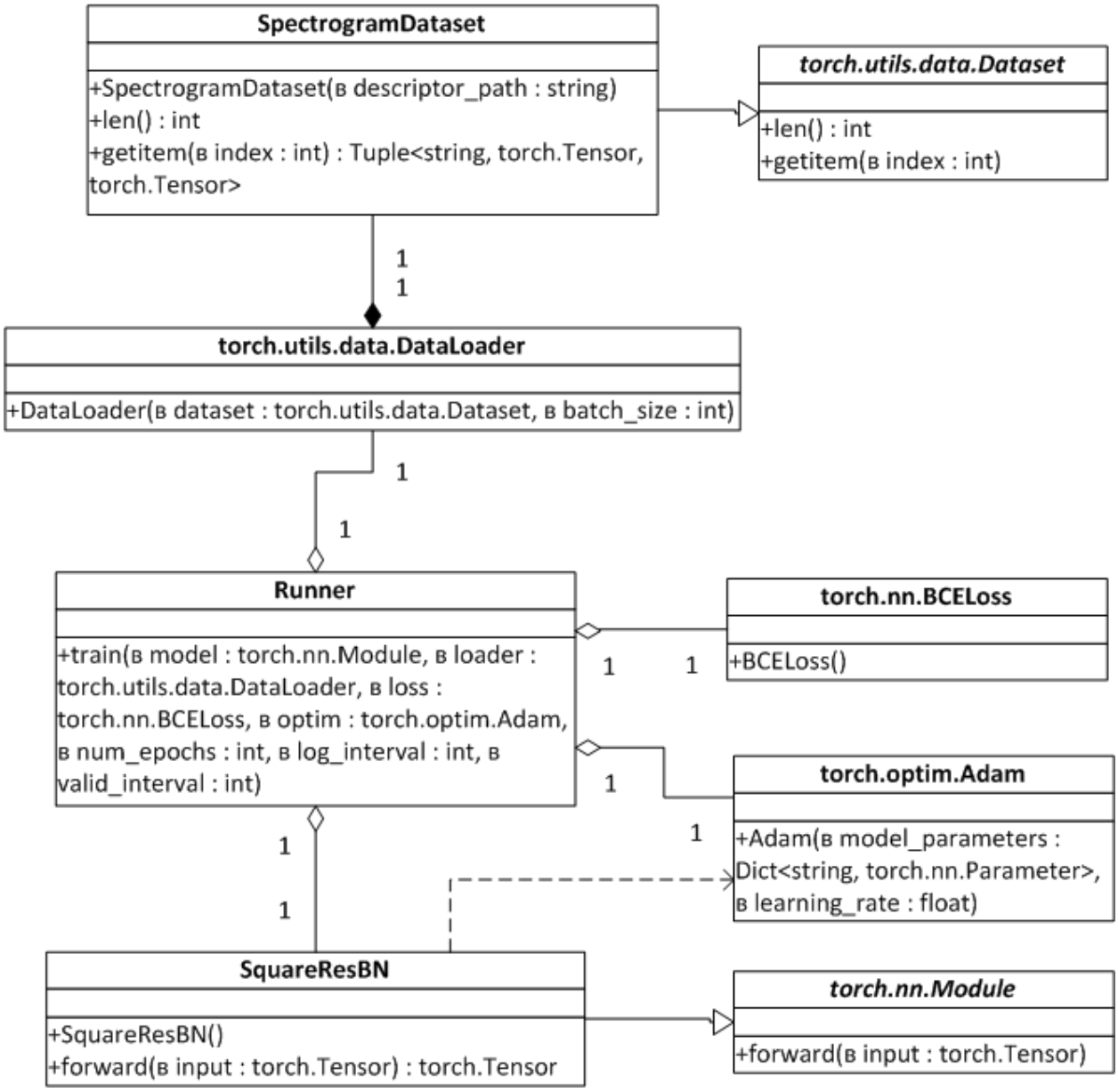


Рисунок 16 – UML диаграмма классов модуля обучения

- модель, которая должна быть реализацией абстрактного класса torch.nn.Module с переопределенным методом forward;
- загрузчики пар спектрограмма и ее метка из соответствующей выборки, которые являются объектами класса torch.utils.data.DataLoader;
- функцию потерь, в качестве которой в данной работе выступает бинарная перекрестная энтропия, реализованная в классе torch.nn.BCELoss;

- объект класса torch.optim.Adam, содержащий реализацию модифицированного метода стохастического градиентного спуска.
- количество эпох обучения;
- интервал логирования значений функции потерь на обучающей выборке;
- интервал валидации.

5.2. Критерии оценки качества работы алгоритмов

В сущности, алгоритм обнаружения речевой активности является бинарным классификатором, отличающим речевую активность от всего остального, поэтому оценить качество работы алгоритмов можно по общеизвестным метрикам, представленным на рисунке 17 [22].

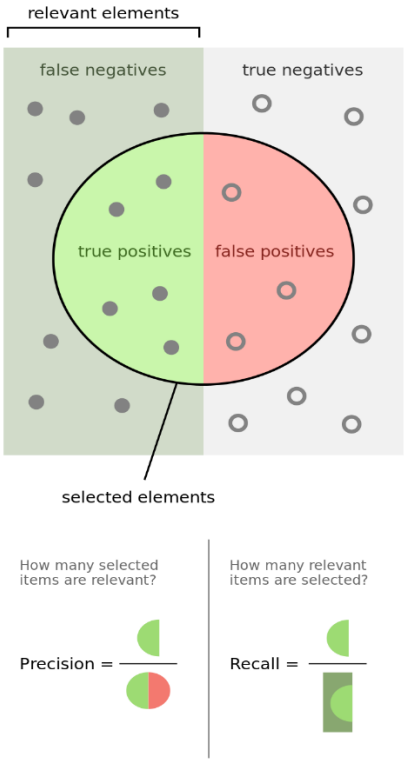


Рисунок 17 – Метрики классификации

Под точностью (Precision) понимают долю объектов (фрагментов аудио), которые модель классифицирует как положительные (речевую

активность) и при этом они действительно являются положительными. Вычисляется как:

$$Pr = \frac{TP}{TP + FP},$$

где TP – истинно положительные, а FP – ложно положительные предсказания.

В то же время полнота (Recall) позволяет определить, какую часть объектов, принадлежащих положительному классу, из всех объектов данного класса модель классифицировала верно. Определяется по следующему выражению:

$$Re = \frac{TP}{TP + FN},$$

где FN – ложно негативные предсказания (пропуск речевой активности).

Очевидно, что чем выше точность и полнота, тем лучше. Однако в реальных задачах высокие точность и полнота одновременно достигаются крайне редко и приходится искать некий баланс. В связи с этим вводят метрику, называемую F1-мера, которая представляет собой гармоническое среднее между точностью и полнотой. Значения определяется по формуле:

$$F_1 = 2 \cdot \frac{Pr \cdot Re}{Pr + Re}.$$

Однако в большинстве случаев предсказания модели не являются бинарными, а лежат в диапазоне значений от 0 до 1. Тогда получаемые величины метрик будут зависеть от порога бинаризации, по которому выход модели относится к одному из двух классов. График зависимости точности от полноты при различных пороговых значениях называется кривая точность-полнота (Precision-Recall curve). Однако сравнивать графики не всегда удобно, поэтому вводят метрику средней точности (Average Precision). Она является площадью под кривой точность-полнота и вычисляется как:

$$AP = \sum_n Pr_n \cdot (Re_n - Re_{n-1}),$$

где Pr_n и Re_n – точность и полнота для n-го порогового значения.

В данной работе оценка качества разрабатываемых моделей осуществляется по средней точности предсказаний на валидационной выборке.

5.3. Эксперименты с архитектурой сверточной нейронной сети

Существует множество различных подходов к построению сверточных нейронных сетей [23, 24]. В данной работе исследование строится на основе архитектуры ResNet [25]. Как и другие современные модели глубокого обучения, данная архитектура состоит из повторяющихся блоков, которые включают в себя несколько типов слоев. На рисунке 18 представлен отличительный элемент топологии ResNet – остаточный блок (Residual Block).

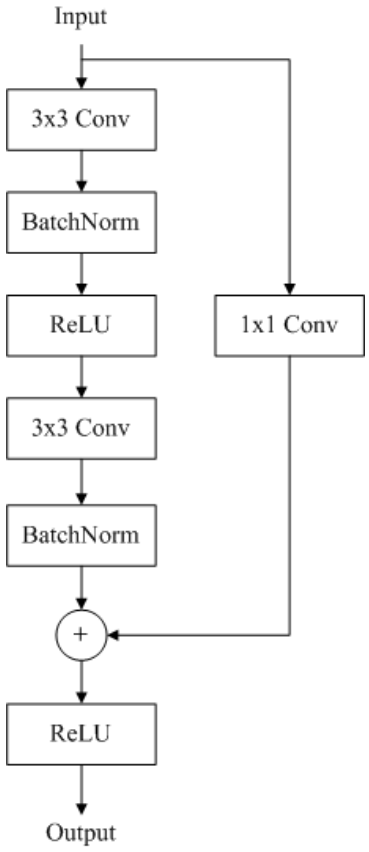


Рисунок 18 – Остаточный блок, предложенный авторами ResNet

Авторы предложили идею обходных соединений (Skip connections), которые позволяют переносить информацию из предыдущего слоя на следующие. За счет этого решается проблема, известная как затухание

градиента (Vanishing gradients) и обучение сверточной нейронной сети ускоряется.

Также улучшить сходимость модели позволяет слой пакетной нормализации (Batch Normalization) [26]. Известно, что обучение глубоких нейронных сетей осложняется тем фактом, что распределения выходов каждого слоя изменяются во время обучения, что вызвано изменением параметров на предыдущих слоях. Это замедляет обучение, требуя меньших значений скорости обучения и тщательной инициализации параметров, а также делает трудным обучение моделей с насыщающимися нелинейностями. Использование слоя пакетной нормализации позволяет постепенно изменять распределения выходов слоя в ходе обучения, что становится возможным благодаря подсчету статистик по пакетам.

Несмотря на достоинства данного блока следует оценить необходимость всех операций, применяемых в нём, для обнаружения речевой активности. Для этого в работе исследуется 4 типа блоков, 3 из которых являются промежуточными версиями остаточного блока ResNet и представлены на рисунке 19.

Данный подход к построению архитектуры с повторением однотипных блоков достаточно популярен, однако необходимо определить из какого количества блоков будет состоять модель.

Как известно, сверточный слой извлекает признаки из локальной области входных данных, размер которой задается ядром свертки. При передаче выходных карт признаков из одного сверточного слоя на вход другого расширяется область входного изображения, на основании которой вычисляется признак. Данная область называется рецептивным полем модели. Оно вычисляется следующим образом [27]:

$$r = \sum_{l=1}^{L} \left((k_l - 1) \prod_{i=1}^{l-1} s_i \right) + 1,$$

где L – количество слоев в сети, k_l – размер ядра свертки l-го слоя, s – шаг ядра свертки на i-ом слое.

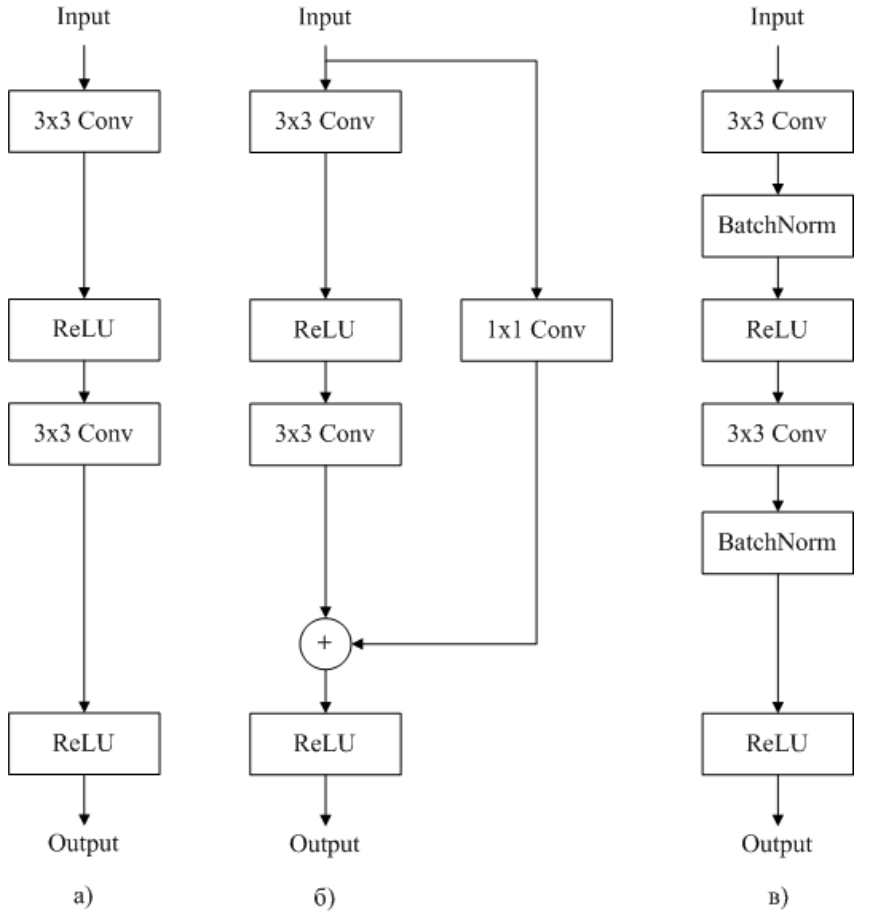


Рисунок 19 – Исследуемые блоки: а) сверточный; б) сверточный с обходным соединением; в) сверточный с пакетной нормализацией

Для достижения компромисса между точностью и вычислительной сложностью модели следует добиться того, чтобы размер рецептивного поля был сопоставим или больше размера входного изображения.

В таблице 1 приведены вычисления рецептивного поля для модели с разным количеством сверточных блоков.

Таблица 1 – Вычисление рецептивного поля

Номер слоя l	Имя слоя	Размер ядра свертки k	Шаг ядра свертки s	Рецептивное поле r
1	Conv	3	1	3
2	MaxPool	2	2	4
3	ConvBlock1.1	3	1	8
4	ConvBlock1.2	3	2	12
5	ConvBlock2.1	3	1	20
6	ConvBlock2.2	3	2	28
7	ConvBlock3.1	3	1	44
8	ConvBlock3.2	3	2	60
9	ConvBlock4.1	3	1	92
10	ConvBlock4.2	3	2	124
11	ConvBlock5.1	3	1	188
12	ConvBlock5.2	3	2	252

Следует отметить, что в данной работе размер входных изображений фиксирован и равен 128 на 128 пикселей. В таком случае можно заключить, что модель с тремя сверточными блоками имеет рецептивное поле, охватывающее только четверть спектрограммы (область 60 на 60 пикселей). В то же время при вычислении признаков, получаемых на выходе сети с четырьмя сверточными блоками, учитывается область 124 на 124 пикселя, то есть практически всё входное изображение. Рецептивное поле модели с пятью

блоками больше входного изображения, что может быть избыточно для решения поставленной задачи.

В данной работе исследуются три архитектуры, состоящие из трёх, четырёх и пяти сверточных блоков на предмет качества классификации и вычислительной сложности. При этом каждая архитектура рассматривается с четырьмя различными типами сверточных блоков.

На данном этапе следует выбрать один из двух основных подходов к предобработке изображений перед подачей их на вход сверточной нейронной сети. При использовании первого подхода производят преобразование диапазона целочисленных значений интенсивности пикселей от 0 до 255 в диапазон дробных значений от 0 до 1. Данный вид предобработки используется, например, в работе [28]. Второй подход включает первый, однако затем следует нормализация данных на основе среднего значения и стандартного отклонения по выборке. Такой вид предобработки встречается при обучении на наборе данных ImageNet [13]. Необходимо сравнить, какой подход позволяет наиболее качественно классифицировать спектрограммы.

Для того, чтобы все эксперименты были сравнимы между собой, следует зафиксировать гиперпараметры обучения и последовательность генератора псевдослучайных чисел. Одним из наиболее важных гиперпараметров является алгоритм оптимизации. В данной работе используется одна из модификаций метода градиентного спуска, известная как Adam [29]. Значение скорости обучения равно 3e-4, размер пакета изображений 256, количество эпох обучения 5. В результате обучения модель с наибольшим значением средней точности (AP) на валидационной выборке сохраняется для дальнейшего использования.

Процесс обучения рассматривается на примере модели с тремя блоками без обходных соединений и пакетной нормализации. На рисунке 20 представлены графики зависимости: а) функции потерь на обучающей выборке от итераций обучения; б) функции потерь на валидационной выборке

от итераций валидации; в) средней точности на валидационной выборке от итераций валидации. При этом исследуются два способа предобработки спектрограммы: синие кривые – преобразование к диапазону от 0 до 1; оранжевые – нормализация.

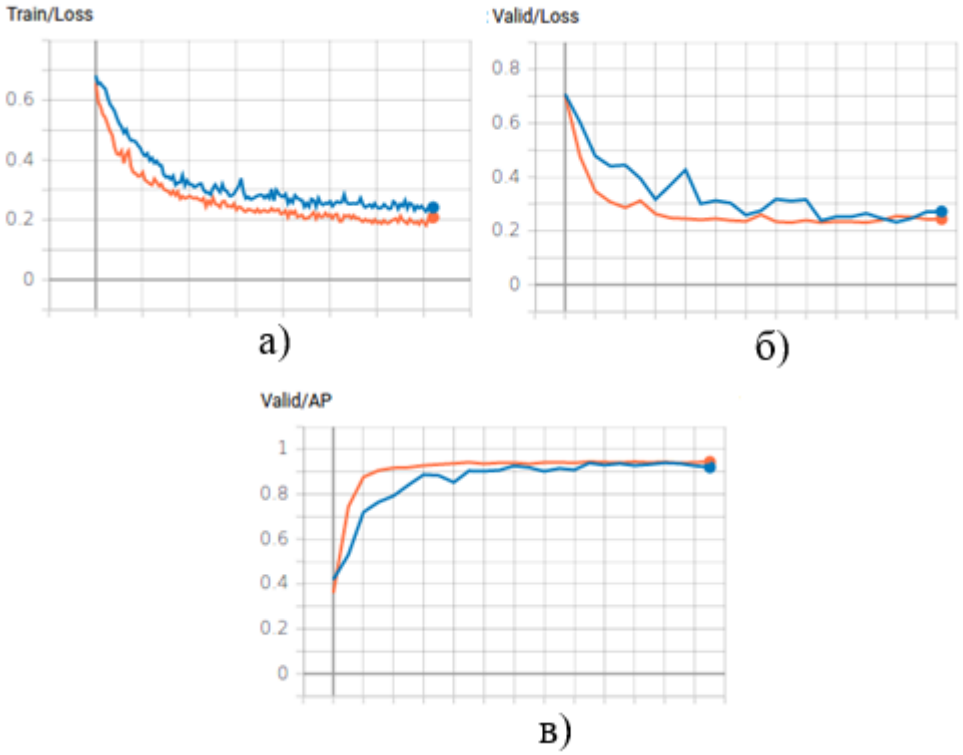


Рисунок 20 – Графики зависимости для модели с тремя блоками без обходных соединений и пакетной нормализации

При анализе графиков можно отметить, что использования второго способа предобработки улучшает сходимость модели. Более выражено данный тезис виден на рисунке 21. На нём представлены те же графики зависимости для модели с тремя блоками с пакетной нормализацией, но без обходных соединений.

Результаты наибольшей средней точности на валидационной выборке для каждой из моделей при различных способах предобработки представлены в таблице 2.

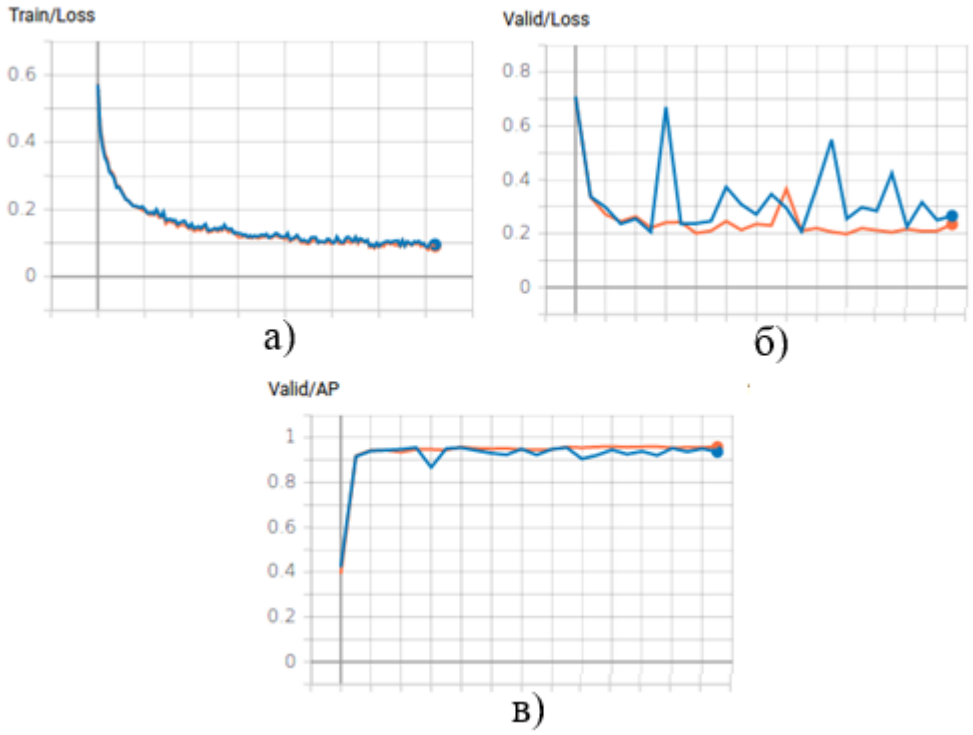


Рисунок 21 – Графики зависимости для модели с тремя блоками с пакетной нормализацией и без обходных соединений

Таблица 2 – Сравнение способов предобработки спектрограммы

Тип блока	Наибольшая средняя точность на валидационной выборке	
	Преобразование к диапазону от 0 до 1	Нормализация входного изображения
Сверточный	0,939	**0,945**
С обходным соединением	0,934	**0,950**
С пакетной нормализацией	0,956	**0,960**
С обходным соединением и пакетной нормализацией	0,960	**0,961**

Проанализировав полученные результаты, был выбран второй способ предобработки, а именно нормализация спектрограммы, так как при использовании данного подхода средняя точность на валидационной выборке всегда выше, чем при преобразовании к диапазону от 0 до 1.

Выбрав способ предобработки, необходимо исследовать модели с четырьмя и пятью сверточными блоками при тех же гиперпараметрах обучения. Результаты наибольшей средней точности на валидационной выборке для моделей с различным количеством и типом блоков приведены в таблице 3.

Таблица 3 – Сравнение моделей с различным количеством блоков

Тип блока	Наибольшая средняя точность на валидационной выборке		
	Количество блоков в модели		
	3	4	5
Сверточный	0,945	0,955	0,966
С обходным соединением	0,950	0,953	0,963
С пакетной нормализацией	0,960	0,961	0,963
С обходным соединением и пакетной нормализацией	**0,961**	**0,965**	**0,968**

По полученным значениям средней точности на валидационной выборке можно заключить, что лучшим качеством обладает модель с пятью остаточными блоками ResNet. Однако, для нахождения компромисса между точностью и вычислительной сложностью следует оценить модели с тремя,

четырьмя и пятью остаточными блоками на тестовых выборках. Результаты исследований приведены в следующем подпункте.

5.4. Сравнение разработанного алгоритма с WebRTC VAD

Устройство алгоритма WebRTC VAD описано в подпункте 2.2. Данный алгоритм функционирует только в четырех режимах, а его предсказания являются бинарными, поэтому нельзя построить по его предсказаниям корректную кривую точность-полнота. Однако можно оценить точность, полноту и F1-меру в каждом режиме отдельно. Результаты оценки качества работы WebRTC VAD на тестовых выборках A и B представлены в таблицах 4 и 5 соответственно.

Таблица 4 – Оценка алгоритма WebRTC VAD на тестовой выборке A

Агрессивность	Точность	Полнота	F1-мера
0	0,424	0,999	0,595
1	0,425	0,999	0,596
2	0,426	0,997	**0,597**
3	0,531	0,602	0,564

Таблица 5 – Оценка алгоритма WebRTC VAD на тестовой выборке B

Агрессивность	Точность	Полнота	F1-мера
0	0,424	0,999	0,595
1	0.425	0.999	0.596
2	0,428	0,998	0,599
3	0,528	0,793	**0,634**

Анализируя полученные результаты, следует отметить, что WebRTC VAD имеет предельно высокую полноту при значениях агрессивности 0, 1 и 2, то есть участки речевой активности практически не пропускаются, однако только порядка 42% участков из общего количества предсказанных являются верными. Иными словами, алгоритм допускает много ложно положительных срабатываний, и его не следует использовать в случаях, когда дальнейшая обработка речевого сигнала является вычислительно затратной.

Несмотря на то, что исследуемые сверточные нейронные сети выбирались по средней точности на валидационной выборке, для сравнимости с аналогом следует вычислить значения точности, полноты и F1-меры на тестовых выборках. Результаты приведены в таблицах 6 и 7.

Таблица 6 – Оценка разработанных моделей на тестовой выборке A

Название архитектуры	Точность	Полнота	F1-мера
СНС с 3 ResNet блоками	0,918	0,921	0,920
СНС с 4 ResNet блоками	0,929	0,944	**0,936**
СНС с 5 ResNet блоками	0,921	0,918	0,919

Таблица 7 – Оценка разработанных моделей на тестовой выборке B

Название архитектуры	Точность	Полнота	F1-мера
СНС с 3 ResNet блоками	0,893	0,912	0,902
СНС с 4 ResNet блоками	0,904	0,932	**0,917**
СНС с 5 ResNet блоками	0,884	0,923	0,903

При сравнении значений F1-меры можно заключить, что разработанный алгоритм обнаруживает речевую активность более

качественно на обеих выборках. Также визуально оценить результаты классификации с учетом скорости работы моделей можно по рисунку 22. При этом следует отметить, что вычисления для WebRTC VAD проводятся на центральном процессоре AMD FX-8320, а для сверточных нейронных сетей на видеокарте Nvidia GeForce GTX 1070.

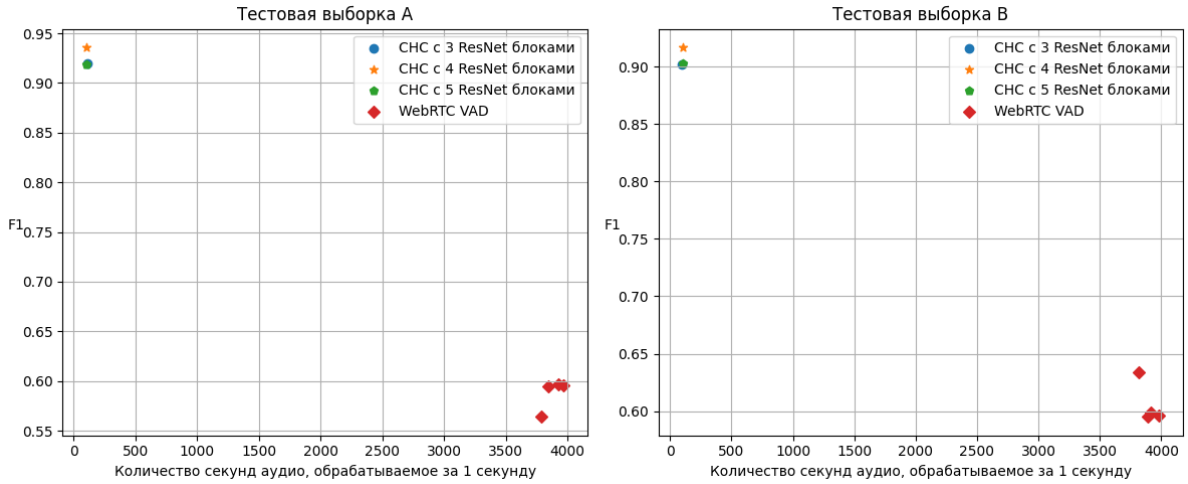


Рисунок 22 – Сравнение разработанных моделей с аналогом

Получившийся результат является удовлетворительным, так как разработанные модели способны за 1 секунду обрабатывать порядка 2 минут аудиозаписи, при этом модель с четырьмя сверточными блоками является наиболее точной. Следует отметить, что только порядка 2 % всего времени занимает прямой проход входных данных через модель, тогда как 98 % времени затрачивается на преобразование аудиозаписей в спектрограммы, расчет которых производится на центральном процессоре.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ВМ83	Теплякову Андрею Борисовичу

Школа	ИШИТР	**Отделение школы (НОЦ)**	ОАР
Уровень образования	Магистратура	**Направление/специальность**	09.04.01 Информатика и вычислительная техника

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	Использовать действующие ценники и договорные цены на потребленные материальные и информационные ресурсы, а также указанную в МУ величину тарифа на эл. энергию
2. *Нормы и нормативы расходования ресурсов*	—
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	Действующие ставки единого социального налога и НДС, ставка дисконтирования = 0,1 (см. МУ)
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	Дать характеристику существующих и потенциальных потребителей (покупателей) результатов ВКР, ожидаемых масштабов их использования
2. *Разработка устава научно-технического проекта*	Разработать проект такого устава в случае, если для реализации результатов ВКР необходимо создание отдельной организации или отдельного структурного подразделения внутри существующей организации
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	Построение плана-графика выполнения ВКР, составление соответствующей сметы затрат, расчет цены результата ВКР.
4. *Определение ресурсной, финансовой, экономической эффективности*	Оценка экономической эффективности использования результатов ВКР, характеристика других видов эффекта
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. *«Портрет» потребителя результатов НТИ*	
2. *Сегментирование рынка*	
3. *Оценка конкурентоспособности технических решений*	
4. *Диаграмма FAST*	
5. *Матрица SWOT*	
6. *График проведения и бюджет НТИ - выполнить*	
7. *Оценка ресурсной, финансовой и экономической эффективности НТИ - выполнить*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН ШБИП	Конотопский В. Ю.	к. э. н.		26.02.2020 г.

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ83	Тепляков А. Б.		

6. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

В работе исследуется алгоритм обнаружения речевой активности в акустическом сигнале для отделения активной человеческой речи от фонового шума или тишины.

К области применения алгоритма можно отнести технологии голосового пользовательского интерфейса: ассистенты, например, Cortana, Siri, Алиса; голосовое управление в автомобиле. Также обработка речевого сигнала может помочь людям с нарушениями слуха. Современные слуховые аппараты усиливают речевой сигнал и подавляют шумовые компоненты.

Целью данного раздела является комплексное описание и анализ финансово-экономических аспектов выполненной работы. Необходимо оценить полные денежные затраты на исследование (проект), а также дать приближенную экономическую оценку результатов ее внедрения. Это в свою очередь позволит с помощью традиционных показателей эффективности инвестиций оценить экономическую целесообразность осуществления работы.

6.1. Планирование и организация работ

При проведении научно-исследовательских работ необходимо организовать группу, способную разработать как сам проект, так и оценить возможные экономические и социальные риски проекта, а также выполнить требования потенциальных потребителей.

В данном пункте представлен полный перечень проводимых работ и их исполнителей. Состав научно-исследовательской группы проекта с представлен в таблице 8.

Перечень этапов и работ в рамках проведения исследования представлен в таблице 9.

Таблица 8 – Состав рабочей группы проекта

ФИО, основное место работы, должность	Роль в проекте
Спицын В. Г., профессор ОИТ ТПУ	Научный руководитель (НР)
Тепляков А. Б., магистрант ОИТ ТПУ	Инженер-программист (ИП)

Таблица 9 – Этапы выполнения научно-исследовательской работы

№	Этапы работы	Исполнители	Вклад исполнителей
1	Составление и утверждение технического задания	НР, ИП	НР – 80% ИП – 20%
2	Поиск материалов по тематике исследования	НР, ИП	НР – 20% ИП – 80%
3	Изучение существующих алгоритмов обнаружения речевой активности, выявление их достоинств и недостатков	НР, ИП	НР – 10% ИП – 90%
4	Календарное планирование работ	НР, ИП	НР – 50% ИП – 50%
5	Разработка структуры алгоритма, решающего поставленную задачу	НР, ИП	НР – 10% ИП – 90%
6	Оценка результатов и итеративное изменение структуры алгоритма	ИП	ИП – 100%
7	Сравнение результатов работы конечного варианта алгоритма с другими решениями в данной области	ИП	ИП – 100%
8	Анализ полученных результатов и подведение итогов проделанной работы	НР, ИП	НР – 20% ИП – 80%
9	Оформление пояснительной записки	НР, ИП	НР – 10% ИП – 90%

6.2. Определение трудоемкости выполнения работ

Трудовые затраты в большинстве случаев образуют основную часть стоимости разработки, поэтому важным моментом является определение трудоемкости работ каждого из участников научного исследования.

Трудоемкость выполнения научного исследования оценивается в человеко-днях и носит вероятностный характер, так как зависит от множества трудно учитываемых факторов. Для определения ожидаемого значения трудоемкости используется следующая формула:

$$t_{\text{ож}\,i} = \frac{3t_{\min i} + 2t_{\max i}}{5},$$

где $t_{\text{ож}\,i}$ – ожидаемая трудоемкость выполнения i-ой работы в человеко-днях; $t_{\min i}$ – минимально возможная трудоемкость выполнения заданной i-ой работы (оптимистическая оценка) в человеко-днях; $t_{max\,i}$ – максимально возможная трудоемкость выполнения заданной i-ой работы (пессимистическая оценка) в человеко-днях.

Теперь необходимо рассчитать длительность этапов в рабочих днях, что можно сделать по формуле:

$$T_{\text{РД}\,i} = t_{\text{ож}\,i} \cdot \frac{K_{\text{Д}}}{K_{\text{ВН}}},$$

где $K_{\text{ВН}}$ – коэффициент выполнения работ, учитывающий влияние внешних факторов на соблюдение предварительно определенных длительностей (принят равным 1); $K_{\text{Д}}$ – коэффициент, учитывающий дополнительное время на компенсацию непредвиденных задержек и согласование работ (принят равным 1,1).

Длительность каждого из этапов работ из рабочих дней следует перевести в календарные дни. Для этого необходимо воспользоваться следующей формулой:

$$T_{\text{КД}\,i} = K_{\text{кал}} \cdot T_{\text{РД}\,i},$$

где $T_{\text{КД}\,i}$ – продолжительность выполнения i-й работы в календарных днях; $K_{\text{кал}}$ – коэффициент календарности.

Коэффициент календарности определяется по следующей формуле [32]:

$$K_{\text{кал}} = \frac{T_{\text{кал}}}{T_{\text{кал}} - T_{\text{вых}} - T_{\text{пр}}} = \frac{366}{366 - 104 - 14} = 1{,}48,$$

где $T_{\text{кал}}$ – количество календарных дней в году; $T_{\text{вых}}$ – количество выходных дней в году (при пятидневной рабочей неделе); $T_{\text{пр}}$ – количество праздничных дней в году.

В таблице 10 приведены расчеты длительности отдельных видов работ и их трудоемкости по исполнителям, занятым на каждом этапе.

Таблица 10 – Расчет трудоемкости выполняемых работ

№ этапа	Трудоемкость работ						Длительность работ в рабочих днях		Длительность работ в календарных днях	
	$t_{\min i}$, чел-дни		$t_{\max i}$, чел-дни		$t_{\text{ож}\,i}$, чел-дни					
	НР	ИП	НР	ИП	НР	ИП	НР	ИП	НР	ИП
1	1	1	4	4	2,2	2,2	2,42	2,42	3,58	3,58
2	2	10	5	15	3,2	12	3,52	13,2	5,21	19,54
3	2	12	5	15	3,2	13,2	3,52	14,52	5,21	21,49
4	1	1	2	2	1,4	1,4	1,54	1,54	2,28	2,28
5	2	10	5	12	3,2	10,8	3,52	11,88	5,21	17,58
6	-	15	-	18	-	16,2	-	17,82	-	26,37
7	-	5	-	6	-	5,4	-	5,94	-	8,79
8	2	3	4	6	2,8	4,2	3,08	4,62	4,56	6,84
9	4	10	5	12	4,4	10,8	4,84	11,88	7,16	17,58
Итого:							**22,44**	**83,82**	**33,21**	**124,05**

Следует отметить, что трудоемкости этапов по исполнителям в рабочих днях далее будут использованы для определения затрат на оплату труда участников.

6.3. Расчет сметы затрат на выполнение проекта

Бюджет научно-технического исследования должен быть основан на достоверном отображении всех видов расходов, связанных выполнением проекта. В процессе формирования бюджета используется следующая группировка затрат по статьям:

- заработная плата исполнителей;
- отчисления во внебюджетные фонды (страховые отчисления);
- расходы на электроэнергию (без освещения);
- амортизация оборудования;
- накладные расходы.

6.3.1. Расчет заработной платы исполнителей

Оплата труда проектной группы, а именно научного руководителя и инженера-программиста рассчитывается на основе трудоемкости выполнения каждого этапа работ $T_{\text{РД}}$, известные из таблицы 10, и месячного оклада исполнителя $З_{\text{мо}}$, величины которых известны из исходных данных.

Среднедневная тарифная заработная плата $З_{\text{дн}}$ рассчитывается по формуле:

$$З_{\text{дн}} = \frac{З_{\text{мо}}}{N_{\text{рд}}},$$

где $N_{\text{рд}}$ – среднее количество рабочих дней в месяце при пятидневной рабочей неделе. В данном случае $N_{\text{рд}}$ следует принять 20.

Также размер заработной платы увеличивается благодаря премиям, дополнительной зарплате за возможное неотработанное рабочее время и

районной надбавке. Величину коэффициента премий $K_{\text{пр}}$ следует принять равной 1,1. Коэффициент дополнительной заработной платы $K_{\text{доп зп}}$ при пятидневной рабочей неделе согласно методическим указаниям [33] равен 1,113. Районный коэффициент $K_{\text{р}}$ для Томской области равен 1,3.

Таким образом, для перехода от тарифной величины заработка исполнителя к соответствующему полному заработку следует первую умножить на интегральный коэффициент $k_{\text{инт}} = K_{\text{пр}} * K_{\text{доп зп}} * K_{\text{р}} = 1{,}1 * 1{,}113 * 1{,}3 = 1{,}591$.

В таблице 11 приведены расчеты затрат на полную заработную плату.

Таблица 11 – Расчеты затрат на полную заработную плату

Исполнитель	$З_{\text{мо}}$, руб./мес.	$З_{\text{дн}}$, руб./день	$T_{\text{РД}}$, дни	$k_{\text{инт}}$	$З_{\text{зп}}$, руб.
НР	47104	2355,2	23	1,591	86183,8
ИП	15700	773,5	84		103373,6
				Итого:	189557,4

6.3.2. Расчет страховых отчислений

В данной статье расходов отражаются обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников. Величина отчислений определяется по формуле:

$$З_{\text{внеб}} = \text{k}_{\text{внеб}} \cdot З_{\text{зп}},$$

где $\text{k}_{\text{внеб}}$ – коэффициент отчислений на уплату во внебюджетные фонды, значение которого в данной работе определено в исходных данных и равно 0,3. Следовательно, страховые отчисления составляют $З_{\text{внеб}} = 189557{,}4 * 0{,}3 = 56867{,}2$ рублей.

6.3.3. Расчет расходов на электроэнергию

В данном пункте необходимо вычислить затраты на электроэнергию, которая потребляется в ходе выполнения проекта на работу оборудования, в частности, персонального компьютера. Затраты рассчитываются по формуле:

$$З_{эл} = P_{об} \cdot t_{об} \cdot Ц_{эл},$$

где $P_{об}$ – мощность, потребляемая оборудованием, кВт; $Ц_{эл}$ – тариф на электроэнергию (определен в методических указаниях); $t_{об}$ – время работы оборудования, часы.

Мощность, потребляемая оборудованием, определяется по формуле:

$$P_{об} = P_{ном} \cdot k_{загр},$$

где $P_{ном}$ – номинальная мощность оборудования, кВт (для персонального компьютера 0,4); $k_{загр}$ – коэффициент загрузки, зависящий от средней степени использования номинальной мощности. Для технологического оборудования малой мощности равен 1.

Время работы оборудования определяется на основе данных таблицы 10 при условии, что продолжительность рабочего дня равна 8 часов по формуле:

$$t_{об} = 8 \cdot T_{РД} \cdot k_t,$$

где k_t – коэффициент использования оборудования по времени, равный отношению времени его работы в процессе выполнения проекта к $T_к$, в данном случае равен 0,8.

В итоге, затраты на электроэнергию $З_{эл} = 0{,}4 \cdot 1 \cdot 8 \cdot 84 \cdot 0{,}8 \cdot 5{,}8 = 1247{,}2$ рублей.

6.3.4. Расчет амортизационных расходов

На данном этапе необходимо оценить амортизацию используемого оборудования, а именно персонального компьютера, за время выполнения проекта. Это можно сделать по формуле:

$$З_{ам} = \frac{Н_А \cdot Ц_{об} \cdot t_{рф} \cdot n}{F_Д},$$

где $Н_А$ – годовая норма амортизации единицы оборудования; $Ц_{об}$ – балансовая стоимость единицы оборудования с учетом ТЗР; $F_Д$ – действительный годовой фонд времени работы соответствующего оборудования для фактического режима его использования в текущем календарном году; $t_{рф}$ – фактическое время работы оборудования в ходе выполнения проекта; n – число задействованных однотипных единиц оборудования.

Для определения $Н_А$ следует обратиться к методическим указаниям [33], содержащим фрагменты из постановления правительства РФ «О классификации основных средств, включенных в амортизационные группы». Оно позволяет получить рамочные значения сроков амортизации (полезного использования) оборудования. Для ПК это 2-3 года. Необходимо задать конкретное значение из указанного интервала, например, 2,5 года. Далее определяется $Н_А$ как величина, обратная сроку полезного использования, то есть $Н_А = 1/2{,}5 = 0{,}4$.

Для определения $Ц_{об}$ при невозможности получить соответствующие данные из бухгалтерии, возможно использовать действующую цену ПК, в данном случае 50000 рублей.

Действительный годовой фонд времени работы $F_Д$ персонального компьютера в текущем году при 248 рабочих днях для пятидневной рабочей недели равен 1984 часам.

Фактическое время работы ПК при восьмичасовом рабочем дне:

$$t_{рф} = 8 \cdot T_{РД} = 8 \cdot 84 = 672 \text{ часа.}$$

В итоге, затраты на амортизацию оборудования составляют:

$$З_{ам} = \frac{0{,}4 \cdot 50000 \cdot 672 \cdot 1}{1984} = 6774{,}2 \text{ рубля.}$$

6.3.5. Расчет накладных расходов

Накладные расходы учитывают прочие затраты при выполнении исследования, не попавшие в предыдущие статьи расходов: печать и ксерокопирование материалов, оплата услуг связи, электроэнергии на освещение и т.д. Накладные расходы определяются по формуле:

$$З_{накл} = k_{накл} * (З_{зп} + З_{внеб} + З_{эл} + З_{ам}),$$

где $k_{накл}$ – коэффициент, учитывающий накладные расходы (определен как 0,1). Тогда:

$$З_{накл} = 0{,}1 * (189557{,}5 + 56867{,}2 + 1247{,}2 + 6774{,}2) = 25444{,}6 \text{ рублей}$$

6.3.6. Формирование бюджета научно-исследовательского проекта

Рассчитанная величина затрат научно-исследовательской работы является основой для формирования бюджета проекта. Определение бюджета на научно-исследовательский проект приведено в таблице 12.

Таблица 12 – Расчет бюджета НТИ

Наименование статьи	Сумма, руб.
Заработная плата исполнителей проекта	189557,5
Отчисления во внебюджетные фонды	56867,2
Расходы на электроэнергию	1247,2
Амортизация оборудования	6774,2
Накладные расходы	25444,6
Итого	**279890,7**

6.4. Оценка экономической эффективности проекта

Анализируя экономическую эффективность исследовательской работы следует отметить, что провести стоимостную оценку результата на данном этапе не представляется возможным. Это обусловлено тем, что алгоритм обнаружения голосовой активности может быть использован в различных

продуктах, например, голосовой ассистент или слуховой аппарат, а рассматривать экономический эффект без четкого определения сферы и границ внедрения некорректно.

Однако в текущей ситуации следует дать систематизированную содержательную характеристику аспектов получаемого эффекта. Во-первых, разработанный алгоритм более точно обнаруживает речевую активность, чем аналог WebRTC VAD (результаты в разделе 5).

Во-вторых, подход, используемый в данном исследовании может быть использован для решения других задач классификации акустических событий, например, при решении задач ежегодного соревнования по обнаружению и классификации акустических сцен и событий DCASE.

В-третьих, разработанное программное обеспечение для формирования набора данных, обучения и тестирования может быть использовано для дальнейшего исследования алгоритмов обнаружения речевой активности с применением сверточных нейронных сетей.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ВМ83	Теплякову Андрею Борисовичу

Школа	ИШИТР	Отделение (НОЦ)	ОАР
Уровень образования	Магистратура	Направление/специальность	09.04.01 «Информатика и вычислительная техника»

Тема ВКР:

Алгоритм обнаружения речевой активности в акустическом сигнале с применением свёрточных нейронных сетей	
Исходные данные к разделу «Социальная ответственность»	
Характеристика объекта исследования и области его применения	В работе исследуется алгоритм обнаружения речевой активности для отделения активной человеческой речи от фонового шума или тишины. К области применения алгоритма можно отнести технологии голосового пользовательского интерфейса.
Перечень вопросов, подлежащих исследованию, проектированию и разработке	
1. Правовые и организационные вопросы обеспечения безопасности 1.1. Специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства. 1.2. Организационные мероприятия при компоновке рабочей зоны.	– ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования»; – ГОСТ 12.2.061-81 «ССБТ. Оборудование производственное. Общие требования безопасности к рабочим местам»; – СаНПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы».
2. Производственная безопасность 2.1. Анализ выявленных вредных и опасных факторов. 2.2. Обоснование мероприятий по снижению воздействия.	Вредные факторы: – недостаточная освещенность рабочей зоны; – повышенный уровень шума на рабочем месте; – отклонение показателей микроклимата в помещении; – повышенный уровень электромагнитных излучений. Опасные факторы: – поражение электрическим током.
3. Экологическая безопасность	– Анализ влияния процесса разработки объекта исследований на окружающую среду. – Обоснование мероприятий по защите окружающей среды.
4. Безопасность в чрезвычайных ситуациях	– Анализ вероятных ЧС при разработке объекта исследований.

	– Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Горбенко М. В.	к. т. н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ83	Тепляков А. Б.		

7. Социальная ответственность

В работе исследуется алгоритм обнаружения речевой активности, который позволяет относить входной акустический сигнал к одному из двух классов: человеческой речи или фоновому шуму.

К области применения алгоритма можно отнести такие технологии голосового пользовательского интерфейса, как ассистенты, например, Cortana, Siri, Алиса, а также голосовое управление в автомобиле. Кроме того, обработка речевого сигнала может помочь людям с нарушениями слуха. Современные слуховые аппараты усиливают речевой сигнал и подавляют шумовые компоненты.

Исследование проводится в 10 корпусе НИ ТПУ в 401 аудитории за рабочим местом, оснащённым персональным компьютером.

Обеспечение производственной и экологической безопасности является необходимым условием реализации как конструкторских, так и исследовательских проектов. В целом, под обеспечением безопасности понимают создание благоприятных рабочих условий для всех лиц, задействованных в работах, предусмотренных проектом, а также условий, обеспечивающих экологическую безопасность окружающей среды.

Первичным этапом в задаче обеспечения безопасности труда является выявление возможных причин потенциальных несчастных случаев, производственных травм, профессиональных заболеваний, аварий и пожаров. Дальнейшими этапами являются разработка мероприятий по устранению выявленных причин и их реализация. Потенциальные причины и риски, а также конкретный набор мероприятий по их устранению, определяются спецификой выполняемых работ и априорными условиями труда (в частности, видом и состоянием рабочих мест исполнителей) [34].

7.1. Правовые и организационные вопросы обеспечения безопасности

7.1.1. Специальные правовые нормы трудового законодательства

В силу того, что разработка и эксплуатация алгоритма обнаружения речевой активности осуществляется за персональным компьютером следует сказать, что для обеспечения безопасности необходимо выполнение правовых норм трудового законодательства по работе с персональным компьютером. Таким образом, рабочее место должно быть организовано с учетом требований:

- ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования»;
- ГОСТ 12.2.061-81 «ССБТ. Оборудование производственное. Общие требования безопасности к рабочим местам»;
- СаНПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы».

В соответствии с государственными стандартами и правовыми нормами обеспечения безопасности предусмотрена рациональная организация труда в течение смены, которая предусматривает:

- длительность рабочей смены не более 8 часов;
- установление двух регламентируемых перерывов (не менее 20 минут после 1-2 часов работы, не менее 30 минут после 2 часов работы);
- обеденный перерыв не менее 40 минут.

Любой, кто разрабатывает или эксплуатирует объект исследования, должен пройти инструктаж по технике безопасности перед началом работ, а также по электробезопасности и охране труда.

7.1.2. Организационные мероприятия при компоновке рабочей зоны

Конструкция рабочей мебели должна обеспечивать возможность индивидуальной регулировки соответственно росту пользователя и создавать удобную позу для работы. Вокруг персонального компьютера должно быть обеспечено свободное пространство не менее 60-120 см.

Работа программиста связана с постоянной работой за компьютером, следовательно, могут возникать проблемы, связанные со зрением. Также неправильная рабочая поза может оказывать негативное влияние на здоровье. Таким образом, неправильная организация рабочего места может послужить причиной нарушения здоровья и появлением психологических расстройств.

Согласно СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы»:

- цветовые параметры должны быть отрегулированы таким образом, чтобы не возникало утомления глаз и головной боли;
- опоры для рук не должны мешать работе на клавиатуре;
- верхний край монитора должен находиться на одном уровне с глазом, нижний – примерно на 20° ниже уровня глаза;
- дисплей должен находиться на расстоянии 45-60 см от глаз;
- локтевой сустав при работе с клавиатурой нужно держать под углом 90°;
- каждые 10 минут нужно отводить взгляд от дисплея примерно на 5- 10 секунд;
- монитор должен иметь антибликовое покрытие;
- работа за компьютером не должна длиться более 6 часов, при этом необходимо каждые 2 часа делать перерывы по 15-20 минут;
- высота стола и рабочего кресла должны быть комфортными.

7.2. Производственная безопасность

7.2.1. Анализ выявленных вредных и опасных факторов

Производственные условия на рабочем месте характеризуются наличием различных опасных и вредных производственных факторов, оказывающих негативное влияние на работников. Под вредными факторами, понимают такие факторы трудового процесса и рабочей среды, которые характеризуются потенциальной опасностью для здоровья, в частности способствуют развитию каких-либо заболеваний, приводят к повышенной утомляемости и снижению работоспособности. При этом вредные факторы проявляются при определенных условиях, таких как интенсивность и длительность воздействия. Опасные производственные факторы способны моментально оказать влияние на здоровье работника: привести к травмам, ожогам или к резкому ухудшению здоровья работников в результате отравления или облучения [35].

При работе за персональным компьютером в течение дня возникают различные производственные факторы, каждый из которых влияет на производительность, работоспособность и физическое состояние.

К вероятным вредным факторам при разработке алгоритма обнаружения речевой активности следует отнести:

- недостаточная освещенность рабочей зоны;
- повышенный уровень шума на рабочем месте;
- отклонение показателей микроклимата в помещении;
- повышенный уровень электромагнитных излучений.

Основным вероятным опасным фактором является поражение электрическим током.

7.2.2. Обоснование мероприятий по защите от воздействия вредных и опасных факторов

Недостаточная освещенность рабочей зоны

Под освещением понимают получение, распределение и использование световой энергии для обеспечения благоприятных условий видения предметов и объектов [34].

В рабочем помещении сочетаются естественное освещение (через окна) и искусственное освещение (использование ламп при недостатке естественного освещения). Светильники в помещении располагаются равномерно по площади потолка, тем самым обеспечивая равномерное освещение рабочих мест.

Зрительные работы программиста относится к разряду III подразряду Г (высокой точности). Параметры искусственного освещения согласно СНиП 23-05-95 «Естественное и искусственное освещение» указаны в таблице 13. Нормы коэффициента пульсации освещенности для III разряда зрительных работ указаны в таблице 14.

Таблица 13 – Нормативные значения освещенности согласно СНиП 23-05-95

Характеристика зрительной работы	Наименьший или эквивалентный размер объекта различения, мм	Разряд зрительной работы	Подразряд зрительной работы	Контраст объекта с фоном	Характеристика фона	Искусственное освещение		
						Освещённость, Лк		
						При комб. освещении		При общ. освещении
						всего	В том числе от общего	
Высокой точности	от 0,3 до 0,5	III	Г	Средний << Большой	Светлый << Средний	400	200	200

Таблица 14 – Нормы коэффициента пульсации освещенности для III разряда зрительных работ согласно СНиП 23-05-2010

Система освещения		Коэффициент пульсации освещенности для III разряда зрительной работы, %
Общее освещение		15
Комбинированное освещение	а) общее	20
	б) местное	15

На данном этапе необходимо оценить параметры освещенности на рабочем месте инженера. Оно имеет следующие характеристики: А = 5 м – длина помещения; В = 4 м – ширина помещения; Н = 3 м – высота помещения; S = 20 м^2 – площадь помещения.

Источником света для общего освещения является светильник для люминесцентных ламп ШОД - 2-80. Количество ламп – 2. Мощность лампы 80 Вт. Длина светильника 1530 мм. Ширина светильника 284 мм. Для создания благоприятных зрительных условий, наименьшая допустимая высота подвеса светильника ШОД над полом 2,5 м.

Размещение светильников в помещении определяется следующими параметрами:

- h_c = 0,2 м – расстояние светильников от перекрытия (свес);
- $h_п = H - h_c = 3 - 0,2 = 2,8$ м – высота светильника над полом, высота подвеса;
- $h_{рп}$ = 0,7 м – высота рабочей поверхности над полом;
- $h = h_п - h_{рп} = 2,8 - 0,7 = 2,1$ м – расчётная высота светильника над рабочей поверхностью.

Уровень необходимого освещения определяется степенью точности зрительных работ. Индекс помещения находится вычисляется как:

$$i = \frac{S}{h * (A + B)} = 0,74$$

Для корректного расчета освещенности необходимо учесть поправки на цвет и степень отражаемости поверхностей. В случае рассматриваемого помещения потолок покрыт белой побелкой, а на стенах бежевые обои. Тогда из методических указаний известно, что коэффициент отражения потолка 0,7, коэффициент отражения стен 0,5.

По индексу помещения и коэффициентам отражения следует определить коэффициент использования светового потока η, который равен 0,36.

Светильники расположены в два ряда, в каждом из которых установлено 2 светильника ШОД - 2-80, учитывая, что в каждом светильнике две лампы белой цветности (ЛБ), мощность которых 80 Вт, а световой поток 5200 лм. Общее число ламп в помещении N = 8. Таким образом, световой поток лампы Ф вычисляется по формуле:

$$\Phi = \frac{E_н \cdot S \cdot K_з \cdot Z}{N \cdot \eta},$$

где $E_Н$ – норма освещенности, 400 Лк; $K_З$ – коэффициент запаса, 1,5 для люминесцентных ламп; Z – коэффициент неравномерности освещения, 1,2 – для люминесцентных ламп.

Значение светового потока равно:

$$\Phi = \frac{400 \cdot 20 \cdot 1{,}5 \cdot 1{,}2}{8 \cdot 0{,}36} = 5000 \text{ лм}.$$

Теперь необходимо осуществить проверку выполнения условия:

$$-10\% \leq \frac{\Phi_{расч} - \Phi_{ном}}{\Phi_{расч}} \cdot 100\% \leq +20\%$$

В данном случае:

$$\frac{5000 - 5200}{5000} \cdot 100\% = -4\%$$

Следовательно, необходимые параметры освещённости обеспечиваются с незначительным отклонением.

Также необходимо так же соблюдать правила подбора шрифта и его фона. В ходе работы, был использован чёрный шрифт и белый фон, размер шрифта – 12-14 пт, что соответствует 0,7 мм для наименьшего объекта различения (точки). Используемые параметры шрифта и фона полностью соответствовали требованиям СНиП 23-05-95 для III разряда зрительных работ.

Повышенный уровень шума на рабочем месте

Под шумом понимают звук, который не несет полезной информации и при определенных обстоятельствах может причинить вред здоровью [34]. Основными источниками шума в помещении являются ПЭВМ, а именно системы охлаждения, расположенные в системном блоке, а также жесткие диски.

При выполнении основной работы на ПЭВМ уровень шума на рабочем месте не должен превышать 50 дБ. Допустимые уровни звукового давления в помещениях для персонала, осуществляющего эксплуатацию ЭВМ при разных значениях частот согласно СН 2.2.4/2.1.8.562-96 «Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой, застройки» приведены в таблице 15.

Таблица 15 – Допустимые уровни звука на рабочем месте согласно СН 2.2.4/2.1.8.562-96

Вид трудовой деятельности	Уровни звукового давления, дБ, в октавных полосах со среднегеометрическими частотами, Гц									Уровни звука (в дБ)
	31,5	63	125	250	500	1000	2000	4000	8000	
Программисты	86	71	61	54	49	45	42	40	38	50

Для снижения уровня шума от персональных компьютеров необходимо регулярно проводить техническое обслуживание, а именно чистить системный блок от пыли, заменять смазывающие вещества, а также возможно применение звукопоглощающих материалов.

Отклонение показателей микроклимата в помещении

Микроклимат производственных помещений – это климат внутренней среды, который определяется действующими на организм человека сочетаниями температуры, влажности и скорости движения воздуха, а также интенсивности теплового излучения от нагретых поверхностей [34].

Оптимальные значения этих характеристик зависят от сезона (холодный, тёплый), а также от категории физической тяжести работы. Для инженера-программиста она является лёгкой (1а), так как работа проводится сидя, без систематических физических нагрузок. Согласно требованиям СанПиН 2.2.4.548-96 «Гигиенические требования к микроклимату производственных помещений», оптимальные параметры микроклимата в офисах приведены в таблице 16.

Таблица 16 – Оптимальные значения характеристик микроклимата по СанПиН 2.2.4.548-96

Период года	Температура воздуха, °С	Относительная влажность, %	Скорость движения воздуха, м/с
Холодный	22-24	40-60	0,1
Тёплый	23-25	40-60	0,1

Если температура воздуха отличается от нормальной, то время пребывания в таком помещении должно быть ограничено в зависимости от категории тяжести работ.

Для создания благоприятных условий труда и повышения производительности, необходимо поддерживать оптимальные параметры

микроклимата производственных помещений. Для этого предусмотрены центральное отопление, вентиляция (искусственная и естественная), искусственное кондиционирование. Все из вышеперечисленных средств используются на рабочем месте.

После фактических измерений были получены следующие параметры: температура воздуха – 23 ℃ (измерено термометром); относительная влажность – 52% (измерено гигрометром); скорость движения воздуха с закрытыми окнами и дверью – 0,1 м/с (измерено анемометром).

Можно отметить, что фактические значения температуры, влажности и скорости движения воздуха соответствуют нормам, установленным в СанПиН 2.2.4.548-96. Для соответствия этим нормам производится проветривание помещения каждый час и ежедневная влажная уборка.

Повышенный уровень электромагнитных излучений

Воздействие электромагнитного излучения на человека зависит от напряженностей электрического и магнитного полей, потока энергии, частоты колебаний, размера облучаемого тела [36]. Работа проводилась на современном компьютере, где значения электромагнитного излучения малы и отвечают требованиям СанПиН 2.2.4.1191-03. «Электромагнитные поля в производственных условиях», которые приведены в таблице 17.

Таблица 17 – Допустимые уровни электромагнитных полей согласно СанПиН 2.2.4.1191-03

Наименование параметров	Допустимые значения
Напряженность электромагнитного поля на расстоянии 50 см вокруг ВДТ по электрической составляющей должна быть не более:	
- в диапазоне частот 5 Гц – 2 кГц	25 В/м
- в диапазоне частот 2 – 400 кГц	2,5 В/м

Плотность магнитного потока должна быть не более: - в диапазоне частот 5 Гц – 2 кГц - в диапазоне частот 2 – 400 кГц	 250 нТл 25 нТл
Напряженность электростатического поля:	20 кВ/м

Основной способ снижения вредного воздействия – увеличение расстояния от источника (не менее 50 см от пользователя). При работе за компьютером специальные экраны и другие средства индивидуальной защиты применены не были.

Электробезопасность

Под электробезопасностью понимают систему организационных и технических мероприятий и средств, обеспечивающих защиту людей от вредного и опасного для жизни воздействия электрического тока, электрической дуги, электромагнитного поля и статического электричества [36]. Персональный компьютер питается от сети 220 В переменного тока с частотой 50 Гц. Помещение, в котором выполнялось исследование, относится к помещениям без повышенной опасности согласно классификации помещений по опасности поражения людей электрическим током [37], так как отсутствуют следующие факторы:

- сырость;
- токопроводящая пыль;
- токопроводящие полы;
- высокая температура;
- возможность одновременного прикосновения человека к имеющим соединение с землей металлоконструкциям зданий, технологическим аппаратам и механизмам, и металлическим корпусам электрооборудования.

Токи статического электричества, наведенные в процессе работы компьютера на корпусах монитора, системного блока и клавиатуры, могут приводить к разрядам при прикосновении к этим элементам. Такие разряды опасности для человека не представляют, но могут привести к выходу из строя компьютера. Для снижения величин токов статического электричества используются нейтрализаторы, местное и общее увлажнение воздуха, использование покрытия полов с антистатической пропиткой.

Короткое замыкание – электрическое соединение двух точек электрической цепи с различными значениями потенциала, не предусмотренное конструкцией устройства и нарушающее его нормальную работу [34]. Короткое замыкание может возникать в результате нарушения изоляции токоведущих элементов или механического соприкосновения неизолированных элементов. Для предотвращения короткого замыкания электрическая цепь содержит автоматический выключатель, который размыкает цепь, если ток превысил допустимое значение.

В качестве мероприятий по предотвращению возможности поражения электрическим током было выполнено следующие:

- введен запрет на работы с задней панелью системного блока при включенном сетевом напряжении;
- исключение перегрева приборов;
- слежение за исправностью электропроводки и целостностью изоляции.

При возникновении неисправностей работы по их устранению производились квалифицированным персоналом.

7.3. Экологическая безопасность

7.3.1. Анализ влияния процесса разработки объекта на окружающую среду

Разработка алгоритма обнаружения речевой активности проводится за персональным компьютером, работа с которым не является экологически

опасной. В данной случае основными воздействиями на окружающую среду являются: повышенное энергопотребление и загрязнением отходами производства и потребления, к которым относятся лампы и макулатура.

7.3.2. Обоснование мероприятий по защите окружающей среды

Вычислительная техника утилизируется с уничтожением информации согласно ГОСТ Р 50739-95 «Средства вычислительной техники. Защита от несанкционированного доступа к информации». Отработанные люминесцентные лампы утилизируются согласно ГОСТ 30772-2001 «Ресурсосбережение. Обращение с отходами. Термины и определения».

В случае выхода из строя они списываются и отправляются на специальный склад, который при необходимости принимает меры по утилизации списанной техники и комплектующих.

Утилизация макулатуры должна производится в соответствии с ГОСТ Р 55090-2012 «Ресурсосбережение. Обращение с отходами. Рекомендации по утилизации отходов бумаги» и ГОСТ 10700-97 «Макулатура бумажная и картонная. Технические условия», согласно которым:

- макулатура должна разделяться на три группы: А – высокого качества, Б – среднего качества, В – низкого качества;
- макулатура каждой группы в зависимости от состава, источников поступления, цвета и способности к роспуску должна соответствовать маркам;
- макулатура должна содержать не более определённого количества примесей, определенного для каждой группы.

Все мероприятия по утилизации выполняются посредством сторонних компаний, имеющих соответствующие лицензии на такую деятельность (согласно ст. 17 Федерального закона от 8 августа 2001 г. № 128 - ФЗ «О лицензировании отдельных видов деятельности»).

7.4. Безопасность в чрезвычайных ситуациях

7.4.1. Анализ вероятных ЧС при разработке объекта исследований

При разработке алгоритма обнаружения речевой активности могут возникать чрезвычайные ситуации различного характера: техногенные, экологические или природные. Однако наиболее типичной ситуацией является возникновение пожара. Данная ЧС может быть вызвана как неисправностью электропроводки, так и нарушением мер пожаробезопасности.

7.4.2. Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС

Пожарная безопасность включает в себя комплекс организационных и технических мероприятий, направленных на обеспечение безопасности людей, предотвращения пожара, ограничение его распространения, а также создание условии для успешного тушения пожара.

Пожарная опасность ПЭВМ обусловлена наличием в применяемом электрооборудовании горючих изоляционных материалов. Горючими являются изоляция обмоток трансформаторов, различных электромагнитов, проводов и кабелей.

Помещение, в котором проводятся исследования, относят к категории Д (пониженная пожароопасность) согласно ППБ-03 [37], характеризующейся отсутствием легковоспламеняющихся веществ и материалов в горячем состоянии. К категории Д относятся помещения, в которых находятся (обращаются) негорючие вещества и материалы в холодном состоянии.

Для того, чтобы избежать возникновения пожара необходимо проводить следующие профилактические работы, направленные на устранение возможных источников возникновения пожара:

- периодическая проверка проводки;

– отключение оборудования при покидании рабочего места;
– проведение инструктажа работников о пожаробезопасности.

Для предотвращения пожара в аудитории с ПЭВМ имеются углекислотный огнетушитель типа ОУ-2 (данный тип огнетушителя подходит для помещений с электрооборудованием (ГОСТ Р 51057-01 и НПБ 155-02), а также пожарная сигнализация ДИП-3СУ (извещатель пожарный, дымовой оптико-электронный точечный).

В случае возникновения пожара необходимо предпринять меры по эвакуации персонала из помещения в соответствии с планом эвакуации (план эвакуации имеется и размещен в помещении).

При отсутствии прямых угроз здоровью и жизни произвести попытку тушения возникшего возгорания огнетушителем. В случае потери контроля над пожаром, необходимо эвакуироваться вслед за сотрудниками по плану эвакуации и ждать приезда специалистов, пожарников. При возникновении пожара должна сработать система пожаротушения, издав предупредительные сигналы, и передав на пункт пожарной станции сигнал о ЧС, в случае если система не сработала, по каким-либо причинам, необходимо самостоятельно произвести вызов пожарной службы по телефону 112, сообщить место возникновения пожара и ожидать приезда специалистов.

Заключение

Подводя итог проведенной работе, хотелось бы отметить, что поставленные задачи были решены. Во-первых, получено представление об области цифровой обработки аудиосигналов, изучены признаки речевой активности, а также методы машинного обучения для её обнаружения. Во-вторых, проведен обзор существующих алгоритмов обнаружения речевой активности, рассмотрены их преимущества и недостатки. В-третьих, выполнен обзор инструментов, с помощью которых возможна разработка детектора речевой активности на основе сверточных нейронных сетей. В-четвертых, спроектированы и реализованы модули для формирования выборок и обучения модели на языке программирования Python. В-пятых, проведено сравнение реализованного алгоритма с WebRTC VAD. Хотя разработанные модели уступают аналогу в скорости обработки аудиозаписей, точность обнаружения речевой активности сверточными нейронными сетями значительно выше на обеих тестовых выборках.

В дальнейшем планируется исследовать другие формы входных данных для модели, в частности, общепринятые в сфере автоматической обработки речи мел-кепстральные коэффициенты [30]. Не менее важным является рассмотрение различных архитектур нейронных сетей, например, рекуррентных моделей.

Также следует отметить, что используемый в данном исследовании подход может быть применен для решения других задач классификации акустических событий, например, в рамках ежегодного соревнования по обнаружению и классификации акустических сцен и событий DCASE [31].

Список публикаций студента

1. Тепляков А. Б. Алгоритм обнаружения речевой активности в акустическом сигнале с применением свёрточных нейронных сетей / А. Б. Тепляков, В. Г. Спицын // Молодежь и современные информационные технологии: сборник трудов XVII Международной научно-практической конференции студентов, аспирантов и молодых ученых, г. Томск, 17-20 февраля 2020 г.: — Томск: Изд-во ТПУ, 2020. — [С. 134-135].
2. Коваль Д. И. Выделение смысловых понятий в медицинских диагнозах при помощи машинного обучения / Д. И. Коваль, И. В. Сушков, А. Б. Тепляков // Молодежь и современные информационные технологии: сборник трудов XVII Международной научно-практической конференции студентов, аспирантов и молодых ученых, г. Томск, 17-20 февраля 2020 г.: — Томск: Изд-во ТПУ, 2020. — [С. 160-161].

Список используемых источников

1. The Best Voice Assistants [Электронный ресурс] / Anne Dennon. – Электрон. текстовые дан. – Режим доступа: https://www.reviews.com/voice-assistant/, (дата обращения: 13.07.2019)
2. Deep Learning Reinvents the Hearing Aid [Электронный ресурс] / DeLiang Wang. – Электрон. журн. – Spectrum IEEE, 2016. – Режим доступа: https://spectrum.ieee.org/consumer-electronics/audiovideo/deep-learning-reinvents-the-hearing-aid, (дата обращения: 14.07.2019)
3. Sound Representation [Электронный ресурс] / Teach Computer Science – Режим доступа: https://teachcomputerscience.com/sound-representation/ (дата обращения: 15.07.2019)
4. J-C Junqua. The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex. Speech Commun. 20(1), 13–22, 1996
5. Cooley, J.W. and Tukey, J.W., An algorithm for the machine calculation of complex Fourier series, Mathematics of Computation, 19(90):297–301, 1965
6. Лайонс Р. Цифровая обработка сигналов: Второе издание. Пер. с англ. – М.: ООО “Бином-Пресс”, 2006 г. - 656 с.: ил.
7. Голосовая биометрия в сфере VoIP [Электронный ресурс] / Олег Тундайкин. – Электрон. журн. – Режим доступа: https://www.it-world.ru/tech/science/142282.html, (дата обращения: 16.07.2019)
8. Moattar, Mohammad & Homayoonpoor, Mahdi. A simple but efficient real-time voice activity detection algorithm. European Signal Processing Conference. 2010
9. Voice Activity Detection for Voice User Interface [Электронный ресурс] / Rudy Baraglia. – Электрон. текстовые дан. – Режим доступа: https://medium.com/linagoralabs/voice-activity-detection-for-voice-user-interface-2d4bb5600ee3 (дата обращения: 18.07.2019)

10. Graf, Simon & Herbig, Tobias & Buck, Markus & Schmidt, Gerhard. Features for voice activity detection: a comparative analysis. EURASIP Journal on Advances in Signal Processing. 2015
11. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. The MIT Press. 2016.
12. Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998. doi: 10.1109/5.726791
13. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) // [Электронный ресурс]. URL: http://www.image-net.org/challenges/LSVRC/ (Дата обращения: 10.10.2019)
14. ITU-T Recommendation G.729 - Annex B: A silence compression scheme for G.729 optimized for terminals conforming to ITU-T Recommendation V.70
15. Python interface to the WebRTC Voice Activity Detector [Электронный ресурс] / GitHub – URL: https://github.com/wiseman/py-webrtcvad (дата обращения: 14.06.2019)
16. Wagner, Johannes & Schiller, Dominik & Seiderer, Andreas & Andre, Elisabeth. (2018). Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant? 147-151. 10.21437/Interspeech.2018-1238.
17. Deep Learning Frameworks Comparison – Tensorflow, PyTorch, Keras, MXNet, The Microsoft Cognitive Toolkit, Caffe, Deeplearning4j, Chainer [Электронный ресурс] / Mateusz Opala. – Электрон. текстовые дан. – Режим доступа: https://www.netguru.com/blog/deep-learning-frameworks-comparison (дата обращения: 12.07.2019)
18. Common Voice [Электронный ресурс] / Mozilla – URL: http://voice.mozilla.org (дата обращения: 18.07.2019)
19. VoxForge [Электронный ресурс] / URL: http://www.voxforge.org/ (дата обращения: 18.07.2019)

20. J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.
21. K. J. Piczak, “ESC: Dataset for environmental sound classification,” in Proceedings of the ACM International Conference on Multimedia. ACM, 2015, in press
22. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves. // Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006
23. Simonyan, Karen and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556, 2015
24. Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv abs/1704.04861, 2017
25. He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778
26. Ioffe, Sergey and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv abs/1502.03167, 2015
27. Computing Receptive Fields of Convolutional Neural Networks [Электронный ресурс] / URL: https://distill.pub/2019/computing-receptive-fields (дата обращения 16.04.2020)
28. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016
29. Kingma, Diederik P. and Jimmy Ba. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980, 2015

30. Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech and Signal Processing, IEEE Transactions on, 28(4):357– 366, 1980.
31. Detection and Classification of Acoustic Scenes and Events [Электронный ресурс] / URL: http://dcase.community (дата обращения: 20.08.2019)
32. Основы функционально-стоимостного анализа: Учебное пособие / Под ред. М. Г. Карпунина и Б. И. Майданчика. - М.: Энергия, 1980. - 175с.
33. Методические указания к выполнению раздела «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» для всех специальностей/ сост. В. Ю. Конотопский; Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2015. – 29 с.
34. Безопасность жизнедеятельности: Учебник для вузов / Под ред. К. З. Ушакова. – М.: Изд-во Московского гос. горного университета, 2000. – 430 с.
35. ГОСТ 12.0.003-74 (с измен. №1, октябрь 1978 г., переиздание 1999 г.) «Классификация вредных и опасных производственных факторов».
36. СанПиН 2.2.4.1191-03. «Электромагнитные поля в производственных условиях».
37. ГОСТ 12.1.009-76 «Электробезопасность. Термины и определения»

Приложение А

Раздел 4
Data sampling module

Студент:

Группа	ФИО	Подпись	Дата
8ВМ83	Тепляков А. Б.		

Консультант Отделения информационных технологий ИШИТР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
профессор ОИТ	Спицын В. Г.	д. т. н.		

Консультант-лингвист отделения иностранных языков ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИЯ	Аксёнова Н. В.	к. филол. н.		

Abstract

Currently, the voice user interface has gained widespread popularity. Such voice assistants as Cortana, Siri, Google Assistant, Alice daily process a significant number of requests, increasing the convenience of performing routine operations by the user [1]. A significant number of both foreign and Russian banks use interactive voice response technology, reducing staff costs. In the automotive industry, hands-free and voice-activated applications allow the driver to interact with people and the car itself while driving, without being distracted from the traffic.

Speech enhancement technologies can also help people with hearing loss. Modern hearing aids enhance the desired speech signal and suppress interfering noise components [2].

Although there are various applications of automatic speech recognition, the developed algorithms face a common problem: it is necessary to detect the presence of speech in an acoustic signal, which is often distorted by noise.

The aim of this work is to develop and implement an algorithm for voice activity detection (VAD) in an input acoustic signal to separate active speech from background noise or silence.

This goal can be divided into the following tasks:

1. search and study articles on research topics;
2. search for existing analogs of the algorithm, consideration of their advantages and disadvantages;
3. overview of tools with which the development of this algorithm is possible;
4. implementation of voice activity detector;
5. comparison of the implemented algorithm with analogues in the accuracy of detection of voice activity.

4. Data sampling module

4.1. Used data sets

The training data set should include the most diverse voice activity of individuals. For these purposes, the open Common Voice and VoxForge data sets were used.

Common Voice is a collection of audio files that contain recordings made by website visitors that read text from a number of public sources, such as blog posts sent by users, old books, movies, and other public data. The data set is used for training and testing automatic speech recognition systems [18].

The VoxForge project was created to collect audio files with speech and for use in open source speech recognition systems such as ISIP, HTK, Julius and Sphinx. The speech corpus is available in the GNU General Public License format [19].

Also, examples of non-speech activity are needed to train the model. For this purpose, the ESC-50 and Urban Noises data sets are used [20].

The Environmental Sound Classification data set is a labeled collection of 2,000 environmental audio records suitable for benchmarking environmental sound classification methods. The data set consists of records lasting 5 seconds, organized in 50 semantic classes (40 examples per class), conditionally divided into 5 main categories: animals; sounds of nature and water; human non-speech sounds; household home sounds; city noises [21].

4.2. Design of the data sampling module

When designing software for data sampling, two main interfaces should be distinguished: for working with data sets and for processing audio files.

Creating the first one allows you to unify the way to get the path to the audio recording in the file system and hide the directory structure specific to each data set.

Processor interface allow implementing specific operations on an audio file, which will be discussed later. At the same time, it should be possible to build a sequence of processors in order to accumulate the results of transformations.

The main processor is the one that implements the operations of forming spectrograms, since in this paper they are considered as input data for convolutional neural networks. As you know, audio recordings can have different durations, so the spectrograms obtained from them are divided into images of the same size using the sliding window method. Details of the process of spectrogram formation are discussed later.

At the moment, one should consider the approach to labeling the resulting images, which is usually one of the most time-consuming stages of preparing the training set. In this case, it is necessary to assign label 1 to each spectrogram if it is known that speech activity is present in this section, otherwise label 0. The following approach to labeling was used in the work. The category to which the spectrogram belongs is determined by the data set from which the audio file was taken.

In the case of labeling data sets with noise, there are no problems, since all spectrograms obtained from the audio signal are labeled as 0. In turn, speech activity is characterized by the presence of silence sections due to pauses between words, the spectrograms of which will be mistakenly labeled as 1. To eliminate this drawback, one implement a processor that first removes silence from the audio recording.

The processors described at this stage are enough to compose a training set from spectrograms with speech and noise activities, however, a significant drawback will be as follows. In the Common Voice and VoxForge data sets, speakers read out various passages in the immediate vicinity of the microphone, whereas in most real-world use cases, the algorithm should be robust to noise exposure. To achieve the goal, audio recordings without silence should be augmented with various noises from the corresponding data sets. The appropriate processor is responsible for implementing the necessary functionality.

Thus, for data sets with speech activity, the processors should be arranged in the following sequence. The first is responsible for removing silence. The following

augments the audio signal. The latter forms spectrograms. For data sets with noise, spectrograms from the audio recording should be produced immediately.

4.3. Data sampling module implementation

The software architecture for data sampling module is presented in the UML class diagram, which is shown in Figure 1.

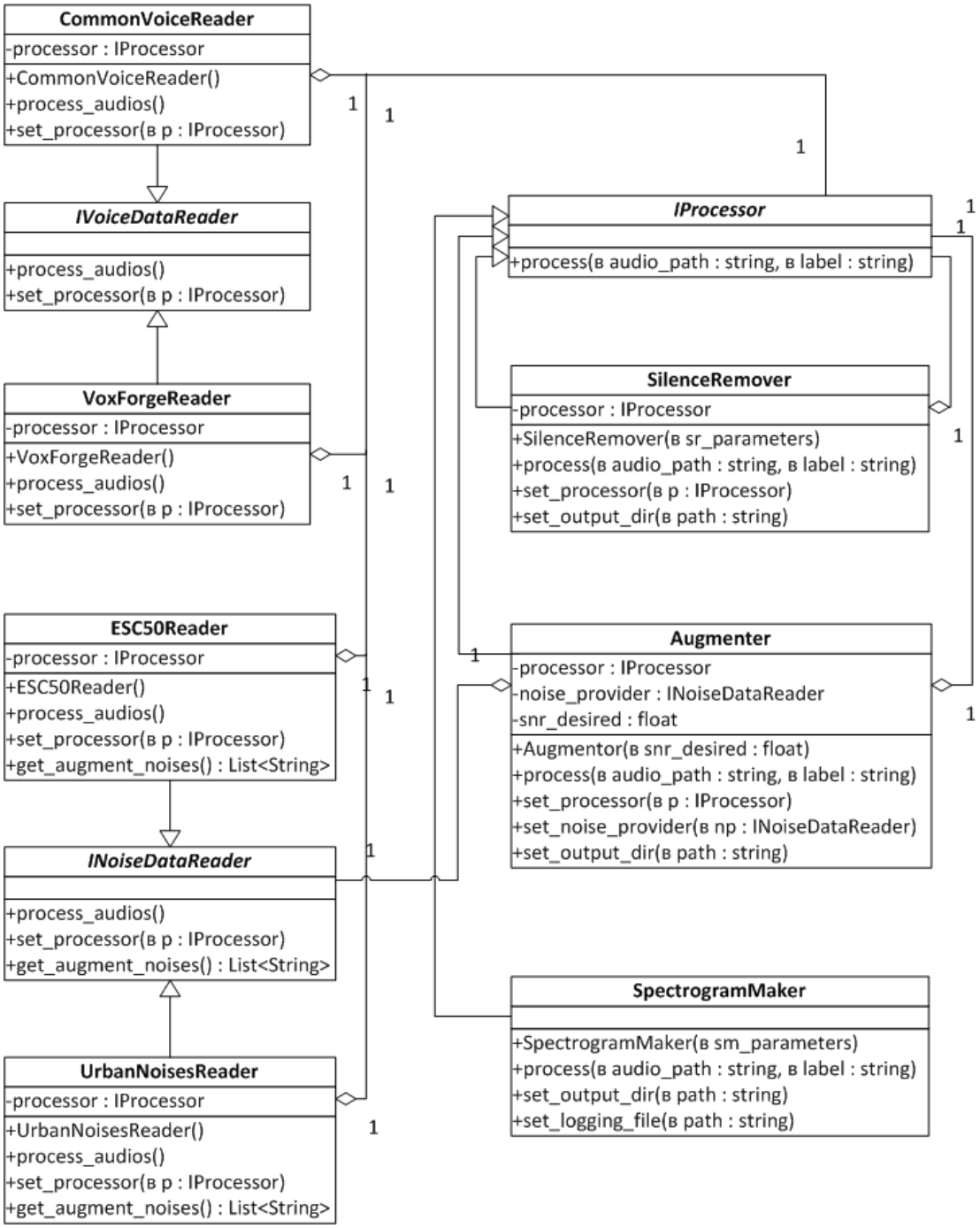


Figure 1 – UML module class diagram

The diagram contains the abstract classes IVoiceDataReader and INoiseDataReader, which provide interfaces for reading datasets with voices and noises, respectively. Their separation is justified by the fact that data sets with noise should transmit audio files for augmentation of voice activity.

The CommonVoiceReader and VoxForgeReader classes implement the IVoiceDataReader interface, namely the methods for setting the processor and handling each of the audio files. In this case, the nature of the processing that will be carried out in the process_audios method is determined by the implementation of the IProcessor interface, and specifically the process method.

In the previous paragraph, three types of processors are defined, which can also be seen in the diagram. Moreover, to create sequentially executed processing operations, the SilenceRemover and Augmentor classes can be set with additional handlers through the set_processor method.

The process of preparing training set begins with the fact that silence is removed from the audio files contained in the data sets with speech activity, for which the SilenceRemover class is responsible. It implements the calculation of the characteristics described in Section 1. It is known from it that such features of speech activity as the short-term energy and the spectrum frequency band ratio can quite successfully remove the silence from an audio signal that was recorded in the absence of extraneous noise. In this case, the threshold values of these features are selected for each speaker when listening.

Then, audio recordings with deleted silence are augmented, for which the Augmentor class is responsible. In this paper, augmentation is understood as the summation of signal amplitudes with speech and noise activities for each sample. This requires four operations.

Firstly, the data member of the Augmentor class noise_provider must be set. It should be an implementation of the INoiseDataReader interface and should return a list of noise audio recordings for augmentation when the get_augmentation_noises

method is called. In this case, the audio files are randomly selected from each category available in the data set.

Secondly, for the correct addition of amplitudes it is necessary that the signals have the same sampling frequency. An audio recording with noise is subject to change, the sampling frequency of which is downsampled to the same value as that of an audio file with speech.

Thirdly, two audio signals most often have different durations, whereas for augmentation of the entire audio file with speech activity they should be equal. Therefore, in case of insufficient duration of noise, it is duplicated the required number of times. If the noise has a longer duration initially or as a result of duplication, then the excess part of the signal is not used.

Fourthly, the amplitudes of both signals most often differ significantly. When they are added, situations are possible when loud noise is superimposed on quiet speech activity. In this case, in reality, the speech will not be heard, but its label will indicate otherwise. There are also possible cases in which the amplitude of the noise signal is much smaller than that of the speech signal, which leads to too little distortion of speech activity. To avoid such situations, it is necessary to control the level of the noise signal. For this, a coefficient k is introduced by which the amplitudes of the noisy signal are multiplied. It is calculated by the formula:

$$k = \sqrt{\frac{SNR}{SNR_{desired}}},$$

where $SNR_{desired}$ – desired value of signal-to-noise ratio as a result of conversion; SNR - initial value of signal-to-noise ratio. It is defined as follows:

$$SNR = \frac{STE(x_V)}{STE(x_N)},$$

where x_V is the amplitude of a signal with speech activity; x_N is the amplitude of a noisy signal; STE is the short-term energy of the signal.

Figures 2 and 3 show examples of audio signal augmentation with initially "weak" and "strong" noises, respectively. Each of the figures in the first row shows the result of the distortion by original noise, while the second row shows the result of distortion by the noise multiplied by the k factor.

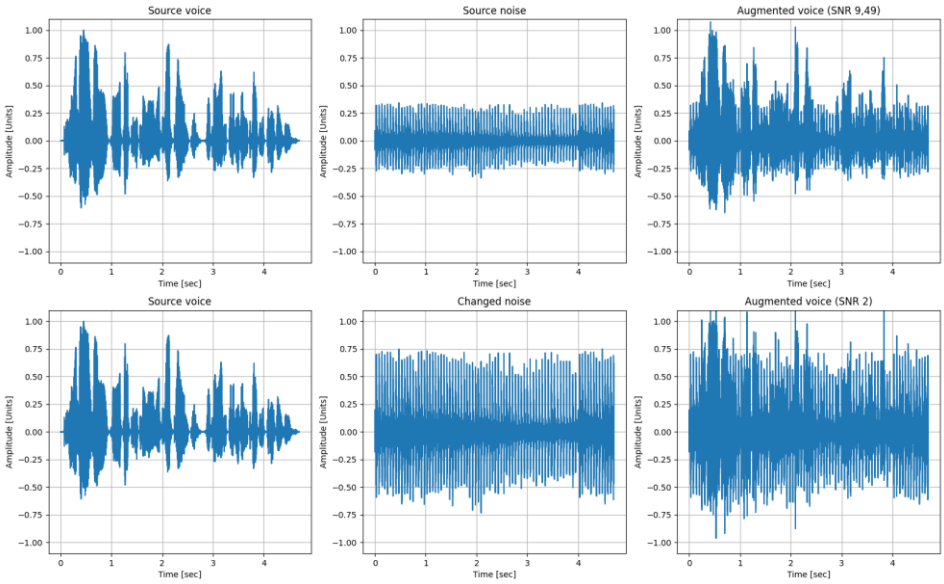


Figure 2 – Example of an augmentation with initially "weak" noise

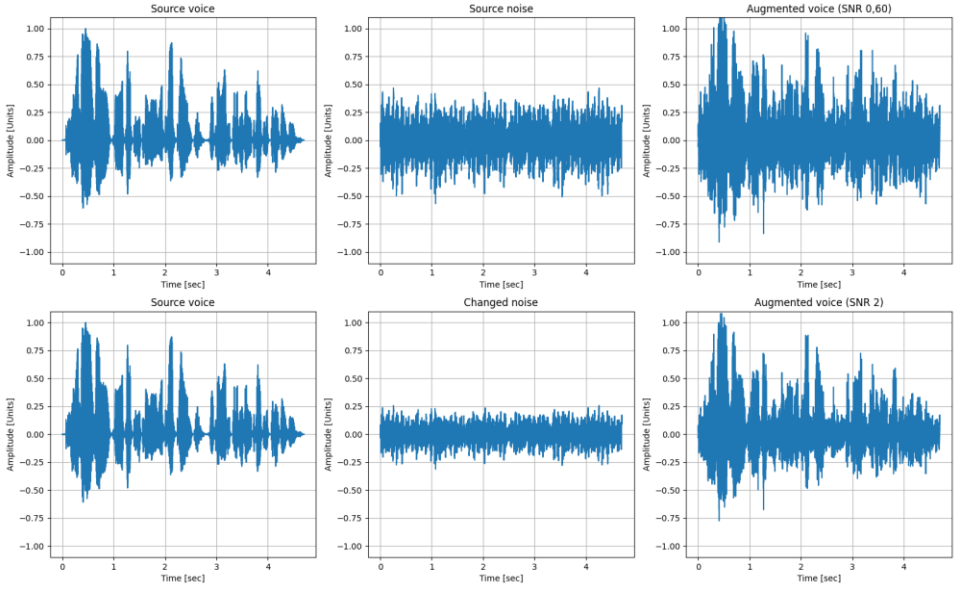


Figure 3 – Example of an augmentation with initially "strong" noise

Thus, the augmentation results in an increased training set size as well as diversity.

The final processor for audio signals with speech activity and the only one for noise is SpectrogramMaker, which implements the following. Initially, each audio file goes through a downsampling to a sample rate of 4000 Hz. Then a spectrogram with the following parameters is formed from the audio signal: window size 400 counts, 396 counts overlapping interval. Thus, on the abscissa axis 1 pixel corresponds to 1 millisecond, and on the ordinate axis frequencies from 0 to 2000 Hz are displayed in 10 Hz steps.

Input for the model is an image in shades of gray 128 by 128 pixels, that is spectrogram parts of the signal with the duration of 128 milliseconds and the power of the signal harmonics at frequencies from 0 to 1270 Hz in 10 Hz steps. In this case, to convert the spectrogram into an image, a normalization takes place, at which the minimum value of the harmonic amplitude is equal to 0 and the maximum value is equal to 255. Examples of spectrograms can be seen in Figure 4.

Figure 4 – Spectrograms for model training (left - speech, middle - drill, right - piano)

4.4. Scenario of using the developed module

To form the training, validation and test sets were used:

- 180 audio recordings from Common Voice;
- 175 audio recordings from VoxForge;
- 1600 audio recordings from ESC-50;
- 4000 audio records from Urban Noises.

It should be noted that categories that may contain speech activity have been removed from noise data sets. As a result of augmentation, the number of speech

activity examples increases by a factor of 8 due to 7 categories from the Urban Noises data set and the use of white noise.

All audio recordings are separated for the preparation of a training, validation and test sets at 70, 10 and 20 percent respectively. In this case, audio recordings related to the test set are aggregated with different values of desired SNR. Therefore, in this paper, the studied models and analogue are tested on two samples: with low interference (the desired SNR is 8) and with high noise activity (the desired SNR is 2). For training and validation sets, the desired SNR is also 2.

As a result of the module operation, 191191 spectrograms for training, 27080 spectrograms for validation, and 53213 spectrograms in each of the two test sets were generated. At the same time the whole data set is quite balanced, as 35 % of the total number of spectrograms belong to the category of speech activity.

Conclusion

Summing up the work carried out, I would like to note that the tasks set have been completed. Firstly, an idea of the area of digital processing of audio signals was received, speech activity features as well as methods of machine learning for its detection were studied. Secondly, the review of existing algorithms of speech activity detection is conducted, their advantages and disadvantages are considered. Thirdly, the review of tools with which it is possible to develop a voice activity detector on the basis of convolution neural networks is made. Fourthly, modules for data sampling and model training in the Python programming language have been designed and implemented. Fifth, the implemented algorithm was compared with WebRTC VAD. Though the developed models are inferior to analogue in speed of processing of audio records, accuracy of detection of speech activity by convolution neural networks is considerably higher on both test samples.

In the future it is planned to study other forms of input data for the model, in particular, the conventional in the field of automatic speech processing Mel Frequency Cepstral Coefficient [30]. It is equally important to consider different neural network architectures, e.g., recurrent models.

It should also be noted that the approach used in this study can be applied to other acoustic event classification problems, such as the annual DCASE acoustic scene detection and classification competition [31].