Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра интеллектуальных информационных технологий

Попандопуло Георгий Петрович

Применение нейросетевых методов для подавления шума в аудиоданных в приложении автоматического распознавания речи

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
к.ф-м.н., ассистент
В.В.Глазкова

Москва, 2022

Аннотация

В рамках данной работы исследуются подходы подавления шума в аудиоданных на основе нейронных сетей. Проведен обзор решений для данной задачи и описаны основные подходы, используемые при их построении. Исследованы существующие наборы данных и метрики качества для оценки получившихся решений. Также описан процесс построения собственных реализации на основе указанных выше методов. Были прежложены собственные подходы решения поставленной задачи. Проведено исследование полученных моделей и приведено их сравнение между собой по основным метрикам качества. Построен демонстрационный стенд, реализующий все исследованные алгоритмы и подходы.

Содержание

1 Введение

1.1 Область применения

Десятилетиями человечество мечтало об описанном в научной фантастике голосовом интерфейсе, с помощью которого герои будущего управляют своими космическими кораблями. В XXI веке выдуманные технологии стали реальностью и в прямом смысле вошли в каждый дом и карман: по дынным за 2017 год от Pew Research [1], 46% американцев используют голосовых помощников, в то время как 62% британцев пользуются услугами голосовых помощников для совершения покупок, проигрывания музыки и поиска в сети.

По оценкам экспертов рынок голосовых ассистентов за последние три года вырос более, чем в 20 раз, что свидетельствует как о заинтересованности потребителей в подобном продукте, так и привлекательности данной сферы для разработчиков программного обеспечения. Таким образом от качества работы помощника будет напрямую зависеть опыт потребителей, а следовательно и его конкурентоспособность.

1.2 Автоматическое распознавание речи

Одной из основополагающих технологий в работе голосовых помощник и ассистентов является технология автоматического распознавания речи (ASR - automatic speech recognition) — именно она отвечает за преобразования речи в текст, который в дальнейшем анализируется программой. По оценкам исследователей из Google [2], одним из важных аспектов, влияющим на качество распознавания речи, является наличие шума в данных. (В таблице 1 приведены значения метрики WER - word error rate, оценивающей процент ошибок при распознавании слов, посчитанной при тестировании 10 различных моделей ASR на наборе данных LibriSpeech)

Таблица 1 — Зависимость WER от качества данных

Используемые данные	WER (%)
Чистые	6.5
Зашумленные	19

Шум — это естественное явление. Он присутствует везде и во всем — на кухне журчит вода из-под крана, на улице шумят автомобили. Сопровождает шум и любую аудиозапись, будь то запись на автоответчике или музыкальные композиции с аудиодиска. Более того, шум может возникать не только из-за наличия посторонних источников звука. Некачественное оборудование записи звука, ошибки при хранение и передачи оцифрованного аудио - все это может послужить причиной помех в аудиозаписи, а значит и повлиять на качество распознавания речи.

1.3 Актуальность

Таким образом, наличие шума в данных - большая проблема для методов автоматического распознавания речи. Предварительная очистка аудиозаписей способна помочь в ситуациях зашумленности входных данных и повысить качество распознавания речи без необходимости изменять/переобучать имеющуюся ASR модель, что позволяет упростить разработку новых сервисов, использующих данную технологию, так и улучшить качество работы уже существующих. Более того, технология подавления шума в аудиоданных может быть использована во множестве других областей - улучшение качества телефонных звонков, голосовых сообщений, а также очистке зашумленных аудиозаписей, невозможных к понимаю без помощи подобных алгоритмов.

2 Постановка задачи

Исследование и реализация нейросетевых методов подавления шума в аудиоданных.

Оценка влияния подавления шума в аудиоданных на качество автоматического распознавания речи на различных наборах данных.

Построение демонстрационного стенда, реализующего весь вышеописанный функционал.

3 Обзор существующих решений

3.1 Цели обзора

Целями данного обзора являются:

1) Поиск существующих наборов данных для задач подавления шума и автоматического распознавания речи.

2) Исследование способов предварительной обработки аудиоданных.

3) Изучение применяемых метрик оценки качества, используемых в данных задачах.

4) Исследование современных методов подавления шума в аудиоданных на основе глубинного обучения.

3.2 Наборы данных

В качестве итоговых рассматриваемых наборов данных были выбраны наиболее популярные наборы, зарекомендовавшие себя в задачах, связанных с обработкой аудиозаписей. По содержанию рассмотренные наборы данных можно разделить на две категории: шумовые данные, «чистые» записи голоса. Также отдельно рассматриваются наборы данных для автоматического распознавания речи.

Такое различие наборов по содержанию не случайно. Во-первых, оно связано с постановкой близких, но все же различных между собой задач. Для подавления шума в аудиоданных нам требуются как сами зашумленные данные, так и их «чистые» варианты, в то время, как распознавание речи не требует отсутствия шумов, однако ему необходима расшифровка того, что на конкретной записи сказано. Наборы данных удовлетворяющие условиям обеих задач существуют, однако по характеристикам (объему, природе представленных данных) они уступают комбинации различных наборов, предназначенных для своих задач. Во-вторых, как уже было сказано, для подавления шума в аудиоданных требуются как сами зашумленные данные, так и их «чистые» варианты, однако, исходя из рассмотренных построенных решений, наиболее эффективным является наложение шума на «чистые» записи отдельно, а не использование готовых зашумлений. При таком подходе стано-

вится возможно существенно расширить обучающую выборку путем наложения случайного выбранного шума, а также снизить эффект «запоминания» моделью конкретной комбинации «речь/шум». Рассмотрим наборы данных из каждой категории подробнее:

3.2.1 Шумовые наборы данных

DEMAND

DEMAND[3] представляет из себя набор различных по своей природе и по силе шумов, записанных в реальных условиях с использованием уникального оборудования - матрицы профессиональных микрофонов (16 штук), установленной на высоте человеческих ушей. В рамках данного набора данных шумы разделены на 6 категорий, каждая из которых состоит еще из 3 подкатегорий, более точно характеризующих окружающую среду во время записи. Описание каждой категории приведено в таблице 2

Таблица 2 — Структура набора DEMAND

Категория	Описание окружающей среды	
Дом	Кухня	на кухне во время приготовления пищи
	Гостиная	в гостиной
	Уборка	ванная комната с работающей стиральной машиной
Офис	Корридор	коридор внутри офисного здания с изредка проходящими работниками
	Совещание	конференц-зал во время обсуждения
	Кабинет	небольшой кабинет с тремя людьми, использующими компьютеры
Общественные места	Столовая	оживленная офисная столовая
	Ресторан	университетский ресторан в обеденное время

Продолжение на след. стр.

Продолжение таблицы 2

	Станция метро	пересадочная зона оживленной станции метро
Транспорт	Автобус	автобус общественного транспорта
	Машина	пассажирское транспортное средство
	Метро	метро
Природа	Спортивная площадка	спортивная площадка с различными видами активного отдыха
	Парк	городской парк
	Река	ручей с проточной водой
Улица	Кафе	терраса кафе на общественной площади
	Площадь	общественная городская площадь с большим количеством туристов
	Пробка	оживленная транспортная развязка

Каждая категория представлена 48 записями (по 16 в каждой подкатегории) длиной 5 минут. Таким образом общая продолжительность данных составляет 24 часа.

Google's AudioSet

Набор данных AudioSet[4] представляет собой масштабную коллекцию размеченных человеком 10-секундных звуковых клипов, взятых из видеороликов YouTube. Структурно все представленные в AudioSet записи разделены на 527 категорий, однако в контексте поставленной задачи нас такое разделение не интересует — с точки зрения построения шумового набора данных представляет интерес отделение записей с речью от всех остальных записей. Таким образом в качестве шумовых данных возможно использовать оставшиеся 526 категорий, которые занимают почти половину всего набора - почти 1.100.000 записей общей продолжительностью около 3.000 часов.

3.2.2 Наборы данных с чистой речью

PTDB-TUG

PTDB-TUG[5] задумывался изначально как набор данных для отслеживания высоты тона, однако наличие в нем студийных, наиболее приближенных к идеальным, записей речи, позволяет использовать его и рамках задачи подавления шума в аудиоданных. В создании данного набора приняло участия 20 носителей английского языка: 10 мужчин и 10 женщин. Каждый испытуемый прочитал 236 из 2342 фонетически насыщенных предложений, причем каждое из них произносилось по крайней мере одним мужчиной и одной женщиной. В общей сложности PTDB-TUG состоит из 4720 записанных высказываний. Общая продолжительность всех записей составляет около 9 часов.

Edinburgh 56 speaker dataset

Edinburgh 56 speaker dataset[6] - англоязычный набор данных, представленный студийными записями носителей языка разных акцентных групп. Существует в двух вариациях: 28 говорящих - 14 мужчин и 14 женщин из одного и того же региона (Англия) и еще 56 говорящих - 28 мужчин и 28 женщин - из разных регионов (Шотландия и Соединенные Штаты). Для каждого человека представлено около 400 проговоренных предложений. Средняя длина записи - 5 секунд, общая продолжительность данных - около 31 часа. Также для каждой записи дополнительная имеется ее текстовая расшифровка.

AVSPEECH

AVSPEECH[7] - это крупномасштабный аудиовизуальный набор данных, содержащий речевые видеоклипы без фоновых шумов. Сегменты длятся 3-10 секунд, и в каждом клипе слышимый звук в саундтреке принадлежит одному говорящему человеку, видимому на видео. В общей сложности набор данных содержит примерно 4700 часов видеосегментов из 290 тысяч видео из YouTube, охватывающих самых разных людей, языки и позы лиц. В рамках задачи подавления шума, визуальная составляющая видеоклипов не играет

никакой роли, однако аудио компонента таких данных оказывается очень ценной, так как представляет из себя записи голоса большого количества обычных людей с отсутствующим или почти отсевающим шумом. Таким образом, извлеченные аудиодорожки из данных AVSPEECH используются для задачи подавления шума в аудиоданных.

3.2.3 Наборы данных для ASR

LibriSpeech

Корпус LibriSpeech[8] составлен на основе аудиокниг, которые являются частью проекта LibriVox, и содержит 1000 часов речи. Для каждого аудио в наборе представлена текстовая расшифровка.

OpenSTT

Корпус OpenSTT (Russian Open Speech To Text (STT/ASR) Dataset)[9] - огромный набор речи на русском языке, собранной из различных доменов. Для каждой аудиозаписи в данном наборе имеется расшифровка, полученная либо вручную, либо с использованием других ASR моделей. Общая продолжительность всех аудиозаписей составляет около 20.000 часов. Данные, по заявлениям его создателей, могут быть использованы для построения решений следующих задач:

1) Распознавание речи;
2) Синтез речи;
3) Устранение шума в аудио;
4) Идентификация голоса;
5) Разделение дикторов;

В таблице 3 представлена структура датасета OpenSTT

Таблица 3 — Структура набора OpenSTT

Источник	Аннотация	Продолжительность (часы)
Радио	Транскрипция	11.996
Публичная речь	Транскрипция	2.709
Youtube	Субтитры	2.117
Книги	Транскрипция/ASR	1.632
Звонки	ASR	819
Другие наборы данных	TTS, начитывание	835

Далее в таблице 4 представлены наиболее важные характеристики всех разобранных наборов данных.

Таблица 4 — Сводка по всем наборам

	Продолжительность (часы)	Язык	Содержание
PTDB-TUG	9ч	Английский	Чистая речь
Edinburgh 56 speaker dataset	31ч	Английский	Чистая речь
AVSPEECH (аудио)	500ч	Английский	Чистая речь
DEMAND	24ч	—	Шум
Google's AudioSet (без речи)	3.000ч	—	Шум
LibriSpeech	1.000ч	Английский	Чистая речь с расшифровками

Продолжение на след. стр.

Продолжение таблицы 4

OpenSTT	20.000ч	Русский	Чистая речь с расшифровками

* В дальнейшем для обозначения пары Edinburgh 56 speaker dataset + DEMAND будем использовать общепринятое название VoiceBank

3.3 Метрики качества

3.3.1 Метрики качества шумоподавления

PESQ

Перцептивная оценка качества речи (PESQ)[10] - это семейство стандартов, включающих методику тестирования для автоматизированной оценки качества речи пользователя телефонной системы. Данный алгоритм представляет собой объективную методику определения качества речевой связи в телефонных системах, которая прогнозирует результаты субъективной оценки качества этого вида связи слушателями-экспертами. Для определения качества передачи речи в PESQ предусмотрено сравнение входного, или эталонного, сигнала с его искаженной версией на выходе системы связи.

SNR

Наиболее распространенной оценкой является соотношение сигнал/-шум (SNR, signal-noise ratio[11]). Этот метод также называют критерием общего отношения сигнал/шум. Он учитывает общее отношение мощности сигнала и шума на всей длительности сигнала. Однако при низкой интенсивности полезного сигнала на каком либо отрезке конечная оценка может быть искажена.

$$SNR = \frac{P_{signal}}{P_{noise}} = \frac{A^2_{signal}}{A^2_{noise}} \tag{1}$$

где P — средняя мощность сигнала, A — среднеквадратичное значение амплитуды сигнала.

Также SNR может быть посчитан в децибелах. В этом случае формула немного видоизменяется:

$$SNR = 10\log(\frac{P_{signal}}{P_{noise}}) = 20\log(\frac{A_{signal}}{A_{noise}}) \quad (2)$$

SSNR

SSNR (segmental signal-noise ratio) является развитием метода соотношения сигнал/шум. В этом случае оценка отношения сигнал/шум производится на интервалах от 15 до 20 мс, что позволяет получить более точную оценку в целом за счёт того, что неравномерная интенсивность сигнала не исказит всей картины в целом.

STOI

Мера разборчивости, которая сильно коррелирует с разборчивостью зашумленных речевых сигналов, например, из-за аддитивного шума, шумоподавления, двоичной маскировки.

CSIG, CBAK и COVL

Данные метрики служат для оценки среднего значения (MOS) субъективных метрик SIG, BAK и OVL соотвественно, которые напрямую оцениваются при опросе респондентов. Для субъективной оценки необходимо указать уровень качества по пятибальной шкале. SIG оценивает искажение входного сигнала, BAK оценивает фоновую интрузивность, а OVL в свою очередь представляет общую оценку качества записи. [10]

3.3.2 Метрики качества распознавания речи

WER

Частота ошибок в словах (WER, Word Error Rate)[12] - это общий показатель качества моделей распознавания речи. Основная трудность измерения качества заключается в том, что распознанная последовательность слов может иметь длину, отличную от эталонной последовательности слов (пред-

положительно правильной). WER основывается на расстояния Левенштейна, работая на уровне слов, а не на уровне фонем. WER является ценным инструментом для сравнения различных моделей, а также для оценки улучшений в рамках одной модели распознавания речи.

$$WER = \frac{S + D + I}{N_1} \tag{3}$$

где S - количество замен, D - количество удалений, I - количество вставок, N_1 - количество слов в оригинале

MER

Частота ошибок в совпадениях (MER, Match Error Rate)[12] - альтернативная, интуитивно более точная мера оценки качества систем распознавания речи. В отличие от WER, MER оценивает вероятность того, что совпадение слов будет неверным.

$$MER = \frac{S + D + I}{H + S + D + I} \tag{4}$$

Где S, D и I несут аналогичный смысл, а H - общее количество совпадений

WIL

Частота потери слов (WIL, Word information lost)[12] передает долю передаваемой (чувствительной к отображению) словесной информации. а коммуникация - это в основном то, для чего предназначена речь, это так что мера является актуальной и показательной.

$$WIL = \frac{H^2}{N_1 N_2} \tag{5}$$

Где H и N_1 несут аналогичный смысл, а N_2- количество слов в гипотезе

3.4 Нейросетевые методы шумоподавления

3.4.1 SEGAN

SEGAN (Speech Enhancement Generative Adversarial Network)[13] является результатом адаптации подходов, используемых в генеративно-состязательных сетях (англ. Generative adversarial network, сокращённо GAN), для задачи подавления шума в аудиоданных. Как и в классических GAN, архитектура SEGAN состоит из двух основных компонент: генеративной модели (G, генератор) и дискриминативной модели (D, дискриминатора). При этом, генератор отвечает за подавление шума, а дискриминатор - за оценивание результата подавления.

Архитектура

Дискриминатор представляет из себя сверточный бинарный классификатор - 11 подряд идущих двумерных сверток с полносвязный слоем в самом конце.

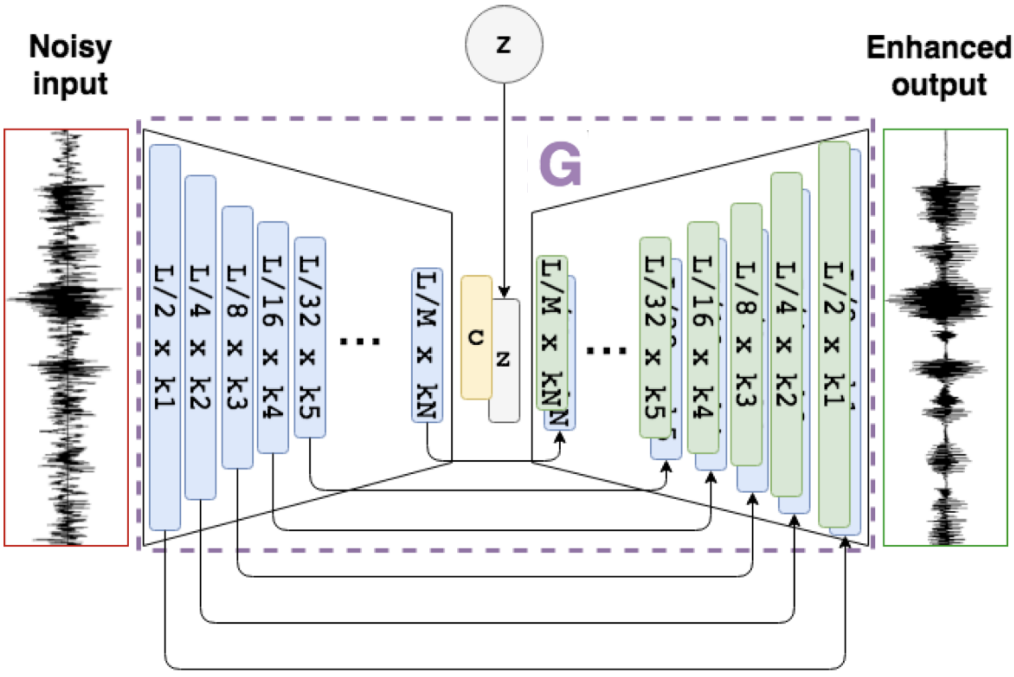

Рисунок 1 — Архитектура генаратора SEGAN

Сложнее устроен генератор. В своей основе он имеет сеть с архитектурой Unet - автоэнкодер с соединениями между слоями энкодера и гомоло-

гичными им слоями декодера (skip connections). Отличием от классического Unet является наличие еще одного входа - прежде чем скрытое представление, полученное в результате работы энкодера, попадет в декодер, с ним конкатенируется вектор, полученный из априорного распределения. Подробнее архитектура генератора представлена на рисунке 1

Обучение

SEGAN обучается точно так же, как и обычная генеративно-состязательная сеть - в 3 этапа: сначала веса обновляются у дискриминатора при входе, предполагающем положительный (True, 1) выход, затем ему на вход подается уже результат работы генератора, однако веса обновляются все еще только у дискриминатор, и наконец дискриминатор замораживается и учится только генератор.

Стоит также отметить особенность подаваемых компонентам обучающих пар. Во-первых, генератор получает зашумленную запись в качестве входа, а чистую - выхода. Также, как было сказано ранее, ему так же требуется вектор из априорного распределения. Во-вторых, дискриминатору также требуются обе записи - зашумленная и чистая. Более того, обе записи подаются ему на вход, а выход вычисляется следующим образом - если для данной зашумлённой записи чистая запись является истинной, то ответ должен быть положительным, в ином случае - отрицательным.

В качества функции потерь предлагается использовать следующую функцию:

$$\begin{aligned} \min_G V_{LSGAN}(G) = \frac{1}{2} E_{z\sim p_z(z),\widetilde{x}\sim p_{data}(\widetilde{x})}[D(G(z,\widetilde{x}),\widetilde{x}) - 1)^2] + \\ +\lambda||G(z,\widetilde{x}) - x||_1 \end{aligned} \tag{6}$$

3.4.2 A Wavenet for Speech Denoising

Wavenet for SD[14] вдохновлен Wavenet - авторегрессионой генеративной моделью для синтеза речи. Это объясняется попыткой отойти от извлечения из аудио спектрограмм для дальнейшей обработки, а работать напрямую

с сырыми данными. Более того, Wavenet for SD сохраняет мощные возможности акустического моделирования Wavenet, значительно снижая ее временную сложность за счет устранения ее авторегрессионного характера.

Архитектура

Wavenet for SD представляет из последовательный набор сверток с разным уровнем дилатации, выходы которых, однако, конкатенируются между собой и пропускаются через пару сверток. Подробнее архитектура модели представлена на рисунке 2.

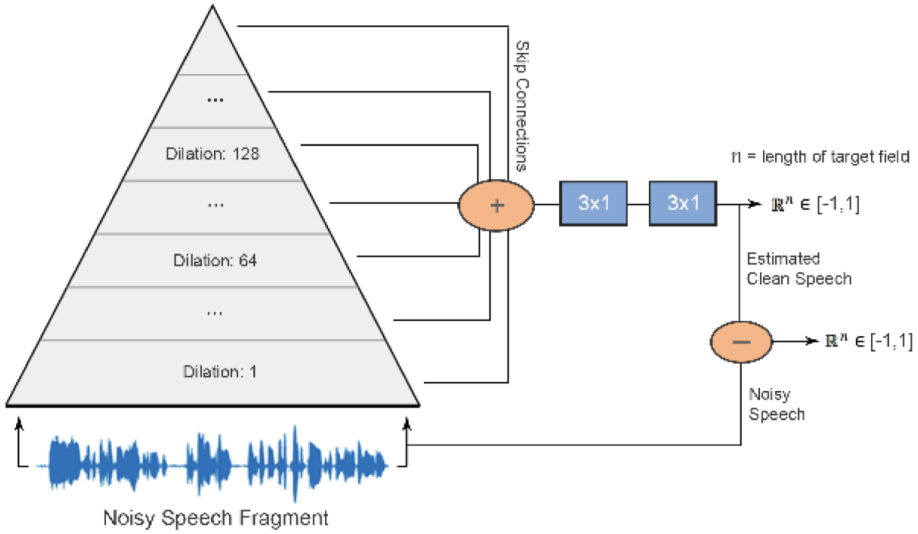

Рисунок 2 — Архитектура Wavenet for SD

Обучение

Модель обучается в supervised режиме, то есть в обучающей паре должны присутствовать как зашумленные данные, так и чистые. Однако, как уже было сказано выше, Wavenet for SD не требует дополнительного извлечения спектрограмм из аудио, а работает с ним напрямую. В качестве функции потерь предлагается использовать функцию сохранения энергии, выглядящую следующим образом:

$$\mathcal{L}(\hat{s}_t) = |s_t - \hat{s}_t| + |b_t - \hat{b}_t| \tag{7}$$

Где s, $\hat{s}$ - ground truth чистый сигнал и предсказанный чистый соответственно, b, $\hat{b}$ - ground truth шум и предсказанный шум соответственно, а $\hat{b} = m - \hat{s}$, где m - зашумлённый сигнал

3.4.3 Listening to Sounds of Silence for Speech Denoising

Listening to Sounds of Silence for Speech Denoising (LSSSD)[15] - подход, вдохновленный неинтеллектуальными методами подавления шума в аудиоданных, а именно извлечение из аудиозаписи участков без голоса, где присутствует только шум (так называемые участки тишины), по которым шум восстанавливается по все длине аудиозаписи. Такой подход позволяет отталкиваться только шума и не зависеть от самой речи, что расширяет сферу его применимости до возможности применять шумоподавления на аудиозаписях с «произвольной» речью, в том числе и на любом языке.

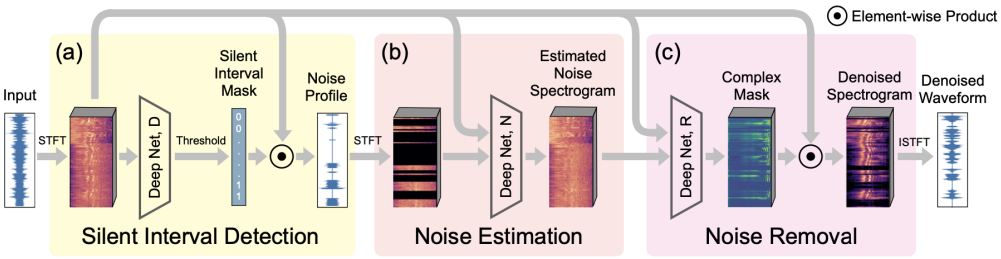

Рисунок 3 — Взаимосвязь компонент сети LSSSD

Архитектура

LSSSD состоит из 3 ключевых компонент: silent interval detection (SID, детектор участков тишины), noise estimation (NE, вычисление шума), noise removal (NR, удаление шума). Их взаимосвязь представлена на рисунке 3.

SID

SID предназначена для обнаружения участков тиишины во входном сигнале. Входным сигналом для этой компоненты является спектрограмма входного (зашумленного) сигнала X . Спектрограмма S_X сначала кодируется двумерным сверточным энкодером в двумерную карту признаков, которая, в

свою очередь, обрабатывается двунаправленным LSTM[16], за которым следуют два полносвязных слоя (FC). Выходом SID является вектор $D(S_X)$. После применения сигмоидной функции, каждое его значение лежит в промежутке [0,1]

Вектор $D(S_X)$ затем расширяется до более длинной маски $m(X)$. Каждый элемент этой маски указывает на уверенность в том, что содержимое конкретного участка входного сигнала - чистый шум. С применением этой маски профиль шума $\widetilde{X}$ оценивается как поэлементным произведением входного сигнала на маску: $\widetilde{X} = X \odot m(X)$.

NE

Профиль шума, полученный в результате работы SID компоненты, представляет из себя не полную картину шума. Однако, поскольку входной сигнал X представляет собой суперпозицию чистого речевого сигнала и шума, наличие полного профиля шума облегчило бы процесс шумоподавления, особенно в присутствии нестационарного шума. Поэтому в рамках данной компоненты оценивается весь профиль шума.

Входы NE компоненты включают как шумный аудиосигнал X, так и неполный профиль шума $\widetilde{X}$. К обоим сигналам применяется STFT (short-time Fourier transfrom, Кратковременное преобразование Фурье) преобразование. Полученные спектрограммы обозначим как S_X и $S_{\widetilde{X}}$ соответственно. Спектрограммы могут быть рассмотрены как 2D-изображения. И поскольку соседние пиксели частоты времени в спектрограмме часто коррелируют, наша цель здесь концептуально сродни задаче рисования изображений в компьютерном зрении. С этой целью мы кодируем S_X и $S_{\widetilde{X}}$ двумя отдельными двумерными сверточными энкодерами в две карты признаков. Затем полученные карты объединяются по каналам и далее декодируются сверточным декодером для оценки полной спектрограммы шума, которую мы обозначаем как $N(S_X,S_{\widetilde{X}})$.

NR

Наконец, мы входной сигнал X от шума. Компонента NR принимает в качестве входных данных как спектрограмму входного сигнала S_X, так и посчитанную ранее полную шумовую спектрограмму $N(S_X,S_{\widetilde{X}})$. Обе спектрограммы обрабатываются двумя отдельными двумерными сверточными энкодерами. Затем получившиеся карты признаков объединяются вместе для передачи в двунаправленный LSTM, за которым следуют три полносвязных слоя. На выходе мы получаем трехмерный тензор, последнее измерение которого определяет действительную и мнимую части маски $c = R(S_X, N(S_X,S_{\widetilde{X}}))$ в частотно-временном пространстве. Другими словами, маска c имеет те же (временные и частотные) размеры, что и S_X.

На заключительном этапе мы вычисляем очищенную спектрограмму S_X^* путем поэлементного умножения исходной спектрограммы S_X и маски c: $S_X^* = S_X \odot c$. Наконец, очищенный аудиосигнал получается путем применения обратного STFT преобразования к S_X^*.

Архитектура вышеописанной сети представлена на рисунке 4.

Обучение

Так же, как и предыдущие методы, LSSSD обучается в supervised режиме, так что нам вновь потребуется как зашумленные данные, так и чистые. Однако, помимо этого, методу требуются еще и шумовая маска

В качестве функции потерь предлагается использовать следующую функцию:

$$\mathcal{L}_0 = E_{x\sim p(x)}[||N(S_x,S_{\widetilde{x}}) - S_n^*||_2 + \beta||S_x \odot R(S_x, N(S_x,S_{\widetilde{x}})) - S_x^*||_2] \quad (8)$$

где N - выход NE компоненты, R - выход NR компоненты, S_X,$S_{\widetilde{X}}$ - спектрограммы водного сигнала и частичного профиль шума соответственно, S_X^*,$S_{\widetilde{X}}^*$ - ground-truth спектрограммы чистой речи и шума соотвественно.

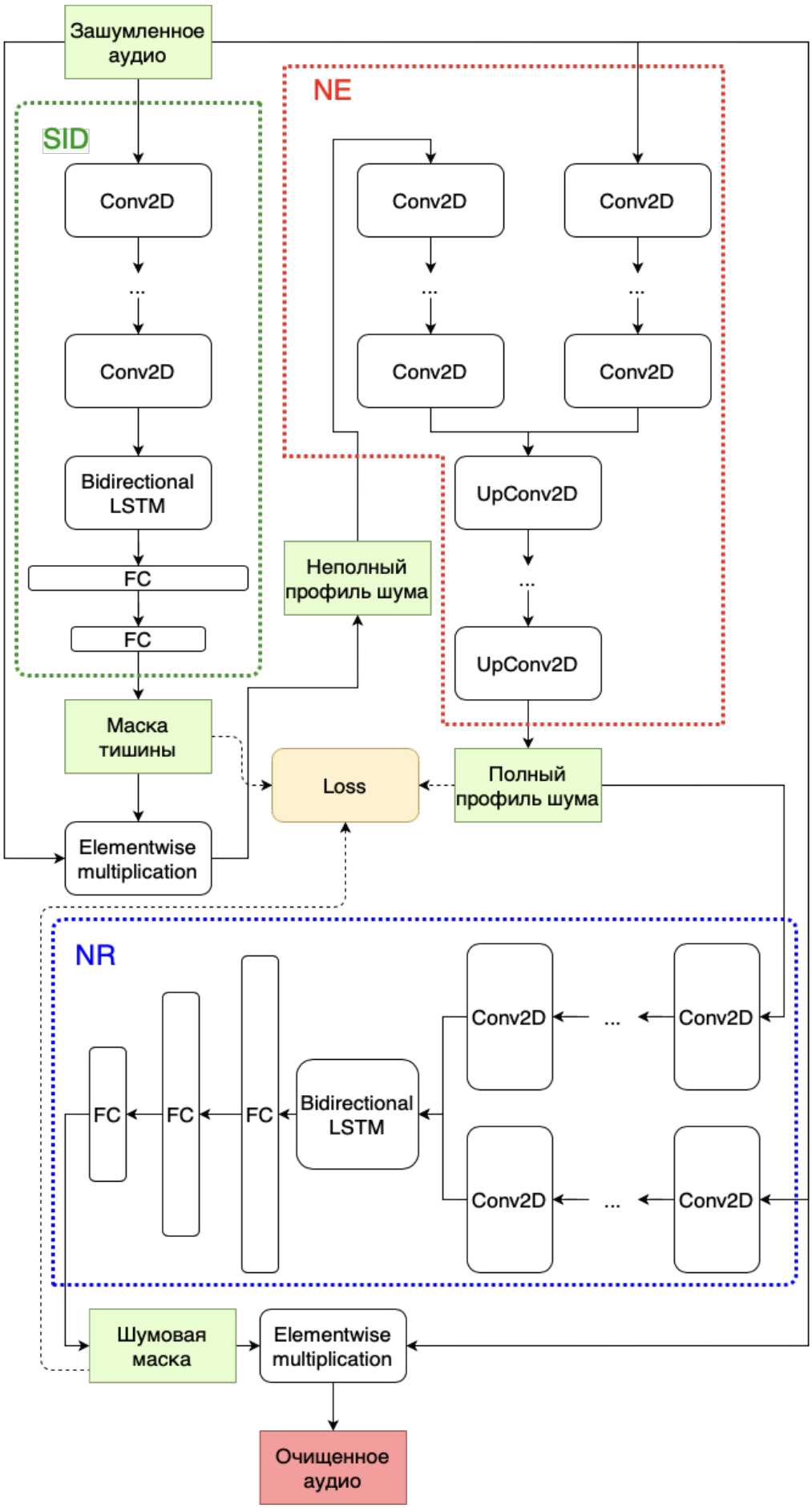

Рисунок 4 — Детальная схема сети LSSSD

Также модель может обучаться в двух режимах - с supervised обучением для SID компоненты и без него. В первом случае она обучается отдельно с применением следующей функции потерь:

$$\mathcal{L}_1 = E_{x\sim p(x)}[l_{BCE}(m(x),m_x^*] \tag{9}$$

где l_{BCE} - бинарная кросс-энтропия[16] $m(x)$ - выход SID компоненты m_x^* - ground-truth маска участков тишины Также тогда необходимо вычислять m_x^* отдельно. Для этого аудио с чистой речью фильтруется по порогу.

3.4.4 Dense Convolutional Recurrent Network

Dense Convolutional Recurrent Network (DCRN)[17] - нейронная сеть, архитектора которой основана на сети UNet[18], была предложена для улучшения качества ASR моделей и повышения их робастности путем одноканального и многоканального улучшения качества речи и подавления в ней шумов.

Архитектура

DCRN состоит из энкодера и декодера и двух LSTM слоев между ним для обработки последовательной природы данных. Выходы из слоев энкодера объединяются с выходами из соответствующих симметричных слоев в декодере (вдоль оси канала). Понижающая размерность в энкодере выполняется с использованием сверток с шагом 2 по частотному измерению, а повышающая размерность в декодере выполняется с использованием субпиксельных сверток (sub-pixel convolution[19]). Кроме того, за пятью слоями в энкодере и пятью слоями в декодере следует «полный» блок. «Полный» блок состоит из пяти сверточных слоев, вход в каждый их которых представляет собой объединение выходов из всех предыдущих слоев в блоке. Количество выходных каналов после каждой свертки в «полном» блоке совпадает с количеством каналов на входе блока. Размеры входа и выхода сети совпадают и равняются - [BatchSize, 2, T, F], где действительная и мнимая части сложены для формирования оси каналов. Все свертки используют фильтры размером 3×3,

за исключением первой и последней, которые используют фильтры размером 5×5. Слои BLSTM используют скрытый слой размера 512 в обоих направлениях. На рисунке 5 представлена архитектура вышеописанной сети.

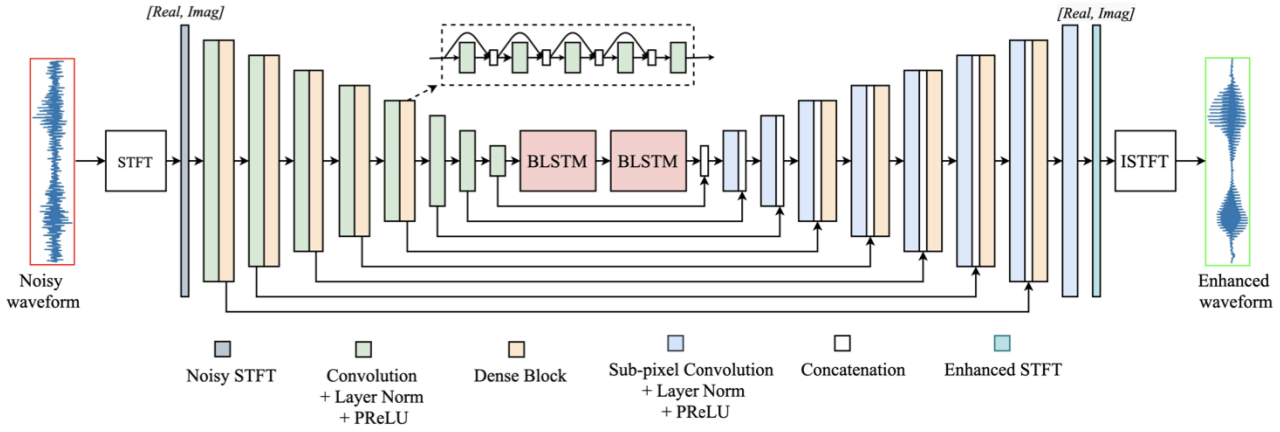

Рисунок 5 — Детальная схема сети DCRN

Обучение

DCRN также является supervised методом, однако, помимо оригинального аудио и зашумленного, никаких других данных для обучения не требуется. Также как и LSSSD, DCRN не работает напрямую с аудиозаписями, предварительно из них необходимо извлечь спектрограммы.

В качестве функции потерь предлагается использовать следующую функцию:

$$L(x,\hat{x}) = \alpha L_t(x,\hat{x}) + (1-\alpha)L_f(x,\hat{x}) \tag{10}$$

$$L_t(x,\hat{x}) = \frac{1}{M}\sum_{n=0}^{M-1}(x[n]-\hat{x}[n])^2 \tag{11}$$

$$L_f(x,\hat{x}) = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}|(|X(t,f)_r| + |X(t,f)_i|) - (|\hat{X}(t,f)_r| + |\hat{X}(t,f)_i|) \tag{12}$$

При этом $x[n]$ и $\hat{x}[n]$ обозначают n-ый элемент чистой и очищенной последовательности соответственно, а M - длину последовательности. $X(t,f)$ и $\hat{X}(t,f)$ же являются T-F единицами спектрограмм x и $\hat{x}$ соответственно. T - количество сэмплов записи, а F - количество частотных ячеек. X_r и X_i,

соответственно, обозначают действительную и мнимую части комплексной переменной X. Параметр α является гиперпараметром, настраиваемым на валидации.

3.4.5 Сравнение методов

Далее в таблицах 5,6 и 7 будут приведены результаты сравнения между собой вышеописанных методов, а также некоторых эвристических подходов. Все тесты проводились в одинаковых условиях с использованием одинаковых данных:

Таблица 5 — Результаты на наборах AVSPEECH + DEMAND

	PESQ	SSNR	STOI	CSIG	CBAK	COVL
Baseline-thres	1,625	6,447	0,737	2,778	2,556	2,168
Spectral Gating[20]	2,542	4,628	0,865	2,819	2,656	2,551
SEGAN	2,227	5,541	0,835	2,594	2,761	2,377
LSSSD	2,795	9,505	0,911	3,659	3,358	3,186
LSSSD + SID	**2,945**	**9,67**	**0,916**	**3,766**	**3,439**	**3,312**

Таблица 6 — Результаты на наборах AVSPEECH + AudioSet

	PESQ	SSNR	STOI	CSIG	CBAK	COVL
Baseline-thres	1,493	4,395	0,685	2,330	2,278	1,867
Spectral Gating	1,845	2,897	0,720	2,065	2,133	1,859
SEGAN	0,942	1,128	0,413	1,137	1,580	1,103
LSSSD	2,304	5,984	0,816	2,913	2,809	2,543
LSSSD + SID	**2,471**	**6,1**	**0,829**	**3,065**	**2,893**	**2,695**

Таблица 7 — Результаты на наборе VoiceBank

	PESQ	STOI	CSIG	CBAK	COVL
SEGAN	2,16	0,93	3,48	2,94	2,80
WaveNet for SD	—	—	3,62	3,24	2,98
LSSSD	**3,16**	**0,98**	**3,96**	**3,54**	**3,53**

3.5 Выводы

Был проведен анализ более 30 научных статей за последние 5 лет. По результатам обзора:

1) Были найдены наборы данных для проведения собственных экспериментальных исследований. Среди рассмотренных наборов будут использоваться:

а) Шумовой набор DEMAND

б) Англоязычные наборы с транскрипциями LibriSpeech и Edinburgh 56 speaker dataset

в) Русскоязычный набор с транскрипциями OpenSTT

2) Были рассмотрены метрики качества как алгоритмов подавления шума в аудиоданных, так и алгоритмов автоматического распознавания речи. Для первой задачи основной выбрана метрика PESQ, для второй - метрика WER

3) Рассмотрены существующие решения поставленной задачи и качество их работы. Наилучший результат показал метод LSSSD - он и будет в дальнейшем использоваться при построение собственного решения. Более того метод DCRN также будет использоваться (подробнее это будет описано в разделе 4.3.3)

4 Исследование и построение решения задачи

В данном разделе будут описаны процесс предварительной обработки данных, реализация выбранных на основе обзора методов, построение собственных подходов на основе проведенного обзора, описание принципов их работы, а также приведены результаты их на основных наборах данных.

4.1 Агрегация данных

На данном этапе для всех наборов данных формируются csv файлы описывающие пути до отдельных файлов набора и хранящую для каждой аудиозаписи ее транскрипции при ее наличии. Данные файлы в дальнейшем служат для описания данных при обучении моделей, а также при тестировании как моделей шумоподавления, так и ASR моделей. Единый формат подобных csv файлов позволяет без особых проблем применить на любом из этапов работы с моделью произвольный набор аудиоданных, в том числе и новых, путем лишь изменения пути до описывающего данные csv файла.

4.2 Подготовка данных

Одним из ограничений при подготовки данных является невозможность моделей шумоподавления работать с аудиозаписями произвольной длины. Поэтому все аудиозаписи нарезаются на записи фиксированной длины - 2 секунды. При этом, нарезание происходит с перекрытием в 1 секунду, то есть из аудиозаписи длиной 3 секунды, мы получаем две аудиозаписи по 2 секунды. В случае, когда изначальная запись длится меньше 2 секунд, то такая запись пропускается. Таким образом, при выбранной частоте дискретизации в 16 кГц каждая запись представляется в программе в виде вектора длиной 32000.

При формировании обучающих и тестовых выборок из всех аудиозаписей выбранного датасета формировался список сэмплов. Элементами данного списка являются следующие четверки - (id, start, end, path) - где id определяет номер сэмпла в списке, start и end - начало и конец текущего сэмпла (относительно исходной аудиозаписи) и path - путь до исходной ауди-

озаписи. В дальнейшем все батчи будут формироваться на основе данного списка сэмплов.

При формировании очередного батча для каждого двухсекудного сэмпла случайным образом выбирается шум из шумового датасета, который на него накладывается. Если же длительность выбранного шума больше 2 секунд, то из него берется случайный двухсекундный отрезок. Далее чистый аудиосэмпл и шумовой сэмпл пропускаются через спектральный биквадратный фильтр (СБФ[21]). СБФ описывается уравнением 13. Такое преобразование позволяет добавить различные акустические эффекты, что увеличивает разнообразие данных. Более того, при каждом смешивании выбирается случайное значение SNR из следующего списка - [-10, -7, -5, -3, -1, 0, 1, 3, 5, 7, 10], определяющее мощность шума относительно мощности выбранного сэмпла. Таким образом, на каждой эпохе итоговые обучающие выборки будут отличаться.

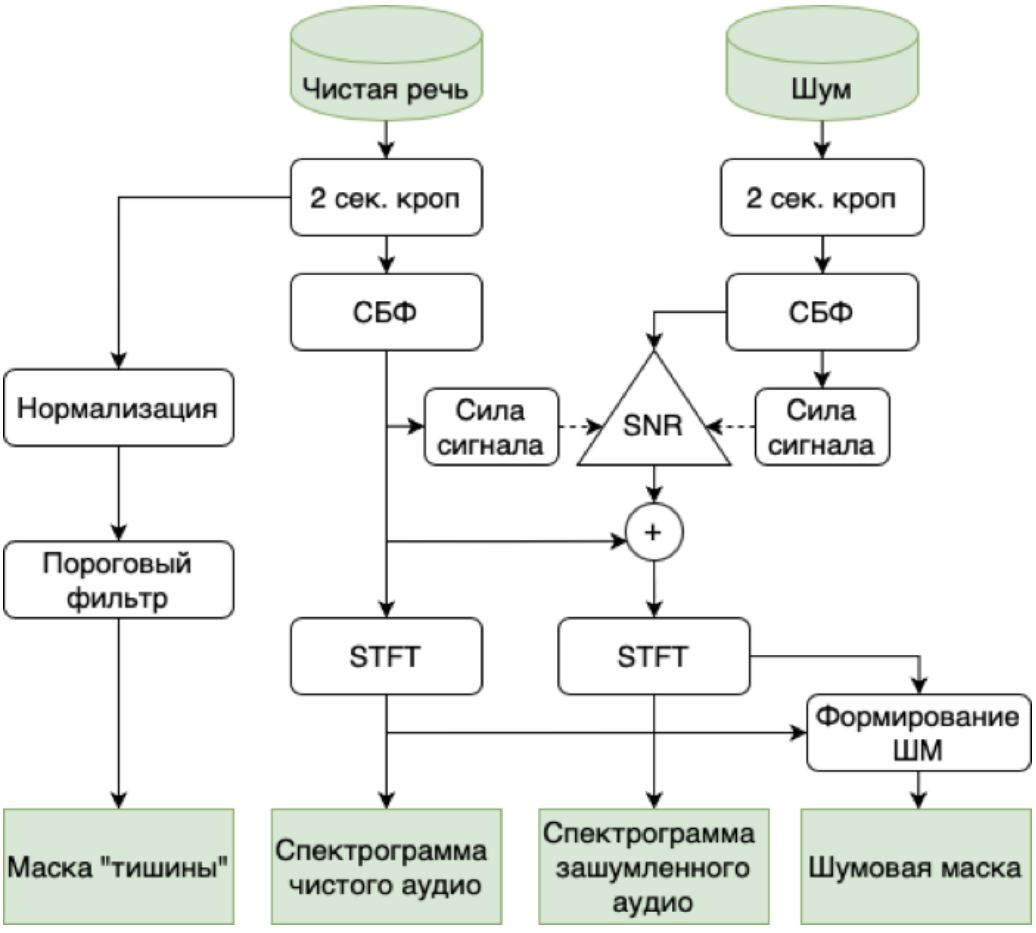

Рисунок 6 — Процесс подготовки данных

$$H(z) = \frac{1 + r_1 z^{-1} + r_2 z^{-2}}{1 + r_3 z^{-1} + r_4 z^{-2}},\ r_i \in Unif(-\frac{3}{8}, \frac{3}{8}) \tag{13}$$

После добавления шума и применения всех вышеописанных трансформаций, к каждой из трех получившихся аудиозаписей (чистая, шумовая и смешанная) применяется кратковременное преобразование Фурье (STFT) со следующими параметрами: n_fft = 510 - определяет количество строк в спектрограмме, то есть количество частотных ячеек, hop_length = 150, win_length = 400 - эти два параметра определяют количество столбцов в спектрограмме, то есть количество сэмплов. По итогу мы получаем 3 спектрограммы, представленные двумерными матрицами размера 256x203, состоящие из комплексных чисел. Однако перед тем, как получившиеся матрицы отправляются в модель, комплексные числа разделяются на действительную и мнимую части, формируя матрицы размера 256x203x2, состоящие уже из действительных чисел. Также на данном этапе формируется маска тишины. Для этого чистая аудиозапись нормализуется (запись делится на максимум модуля записи), а затем разбивается на блоки размера 534 (в случае неполного последнего блока, он дополняется средними значениям имеющихся в последнем блоке элементов), на основе которых формируется новый вектор, состоящий из средних значений по выделенным блокам. Далее получившийся вектор фильтруется по порогу. В случае значения, превосходящего порог, данный участок помечается как речь, в обратном - как тишина. В качестве порога экспериментальным путем было выбрано значение 0.07. Таким образом мы получаем вектор длиной 60, каждое из чисел в котором характеризует наличие речи на участке продолжительностью $2/60 = 1/30$с. На рисунках 6 и 7 представлены общий процесс подготовки данных и извлечения признаков и пример маски тишины соответственно.

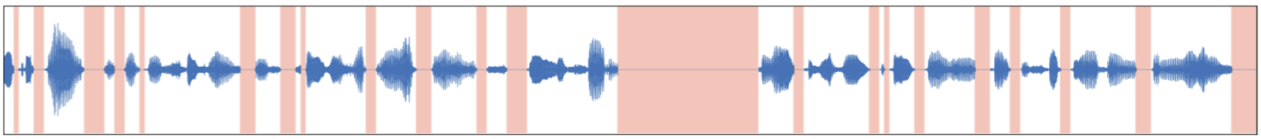

Рисунок 7 — Маска тишины

4.3 Модели шумоподавления

Далее будут описаны все построенные модели шумоподавления.

4.3.1 LSSSD

Данная модель является имплементацией предложенной архитектуры из соответствующей статьи. При этом SID компонента не обучается отдельно.

4.3.2 LSSSD frozen

Данная модель с точки зрения архитектуры совпадает с предыдущей, однако был изменен подход в обучение с точки зрения функции потерь (loss-функции) - в изначальной модели (LSSSD) веса по всем компонентам изменялись по совместному значению функции потерь (сумме функции потерь по компонентам), что приводило к уменьшению общему значения loss-функции, однако потери по отдельным компонентам увеличивалась. Данная проблема была решена путем обучения каждой компоненты совместно, но независимо друг от друга, то есть веса текущей компоненты меняются только по функции потерь данной компоненты.

4.3.3 LSSSD + DCRN

Данная модель является объединением моделей из обзора. С точки зрения архитектуры DCRN выступает в роли дополнительной компоненты. При этом изначальные три компоненты модели LSSSD обучаются совместно (как в изначальном варианте), а DCRN независимо. Функция потерь модели DCRN также претерпела изменения - она была урезана до составляющей. На рисунке 8 представлена архитектура предложенной модели.

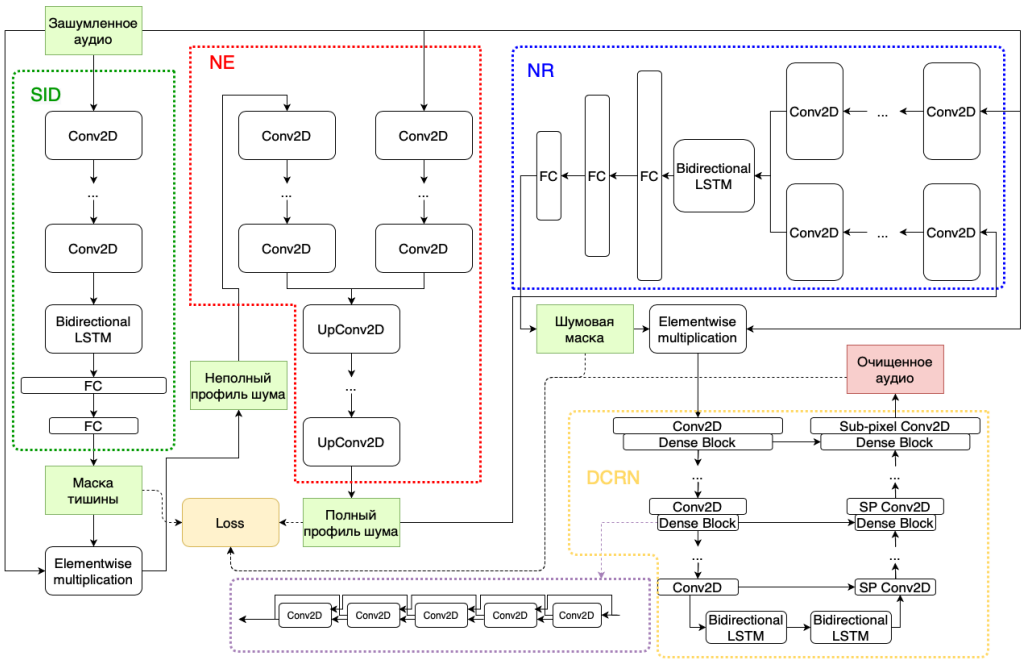

Рисунок 8 — Архитектура сети LSSSD+DCRN

4.4 Модели ASR

В качестве подхода автоматического распознавания речи использовалась предобученная модель на основе тулкита ESPNet[22]. ESPNet в основном фокусируется на end-to-end автоматическом распознавании речи (ASR) и использует популярные нейросетвые библиотеки - Chainer[23] и PyTorch[24] в качестве основного механизма глубокого обучения. ESPnet также следует стилю тулкита Kaldi ASR для обработки данных, извлечения функций/форматирования, что обеспечивает полную настройку моделей для распознавания речи и других задач по обработке речи.

Модель выбиралась из открытого источника Zenodo[25]. В результате, для английского языка была выбрана модель [26], обученная на датасете LibriSpeech, а для русского - модель [27], обученная на датасете OpenSTT.

4.5 Выводы

По результатам исследований:

1) • Был предложен метод агрегации данных, а также рассмотрены методы их предварительной обработки и аугментации; • Были подробно рассмотрены методы шумоподавления в аудиоданных, выявленные во время обзора

(методы ?? и ??, а также предложены модификации, построенные на их основе • Были рассмотрены сущетсвующие предобученные модели автоматического распознавания речи для русского и английского языка;

Для проверки и обоснования достоверности исследований был создан экспериментальный стенд, описанный в разделе 5.1, и проведены эксперименты, описанные в разделе 5.2.

5 Описание практической части

5.1 Программная реализация

В практической части работы реализован автоматизированный компонентный экспериментальный стенд, архитектура которого представлена на рисунке 9, на языке программирования Python 3 с использованием ряда библиотек с открытым исходным кодом, основными из которых являются следующие:

1) Pandas v1.0.0: для агригации данных и результатов;
2) NumPy v1.18.1: для эффективной работы с матричными вычислениями;
3) Librosa v0.9.1: для обработки аудио файлов;
4) Matplotlib v3.1.3: для визуализации данных двумерной графикой;
5) PyTorch v1.11.0: для решения задач построения и обучения нейронных сетей.

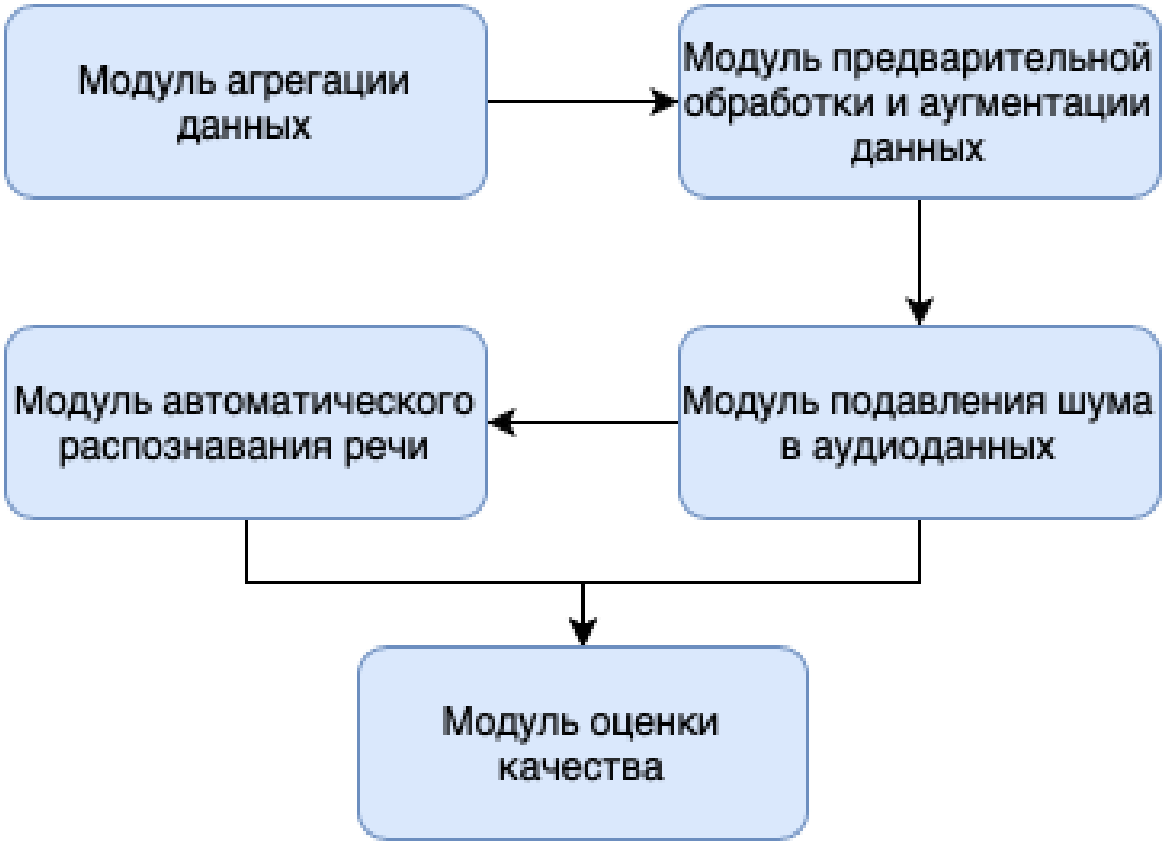

Рисунок 9 — Архитектура экспериментального стенда

5.1.1 Модуль агрегации данных

В рамках данного модуля происходит унифицирование наборов данных, описанных в разделе 3.2, в соответствии со структурой, заданной в разделе 4.1. Файлы с описанием наборов данных сохраняются для следующего этапа конвейера.

5.1.2 Модуль предварительной обработки и аугментации данных

На данном этапе, в соответсвии со схемой 6 и разделом 4.2, происходит предварительная обработка данных

5.1.3 Модуль подавления шума в аудиоданных

Далее к подготовленным на прошлом шаге аудиозаписям применяются различные модели шумоподавления, описанные в разделе 4.3. Более того, для каждой из модели доступна тонкая настройка засчет изменения параметров через конфигурационный файл.

5.1.4 Модуль автоматического распознавания речи

На следующем этапе экспериментального стенда тройки аудиозаписей, состоящие из оригинальной записи, зашумленной и очищенной, обрабатываются одной их двух моделей автоматического распознавания речи, описанных в разделе 4.4.

5.1.5 Модуль оценки качества работы

В заключительной части происходит оценка качества работы на тестовых наборах как моделей шумоподавления, так и моделей автоматического распознавания речи. Проводится расчет метрик оценки качества шумоподавления, описанных в разделе 3.3.1, а также метрик оценки качества распознвания речи из раздела 3.3.2

5.2 Экспериментальные исследования

Было проведено несколько экспериментов по обучению моделей с предложенными архитектурами при различных конфигурациях. Для обучения моделей использовался набор VoiceBank, разделенный на тренировочные и валидационные выборки в соотношении 9/1. В качестве оптимизатора был выбран Adam[28] с уменьшающейся скоростью обучения. Описание экспериментов представлено в таблице 8.

Таблица 8 — Описание экспериментов

Модель	Loss - функция	Количество эпох	Комментарий
LSSSD	$\text{SID}_{loss} + NR_{loss} + NE_{loss}$	200	Базовая модель из статьи
LSSSD frozen	$\text{SID}_{loss} + NR_{loss} + NE_{loss}$	200	Компоненты модели обучаются независимо
LSSSD + DCRN	$\text{SID}_{loss} + NR_{loss} + NE_{loss} + Lt_{loss}$	200	DCRN используется как четвертая компонента

На рисунке 10 изображен график с динамикой значений функций потерь на валидации соответствующих моделей.

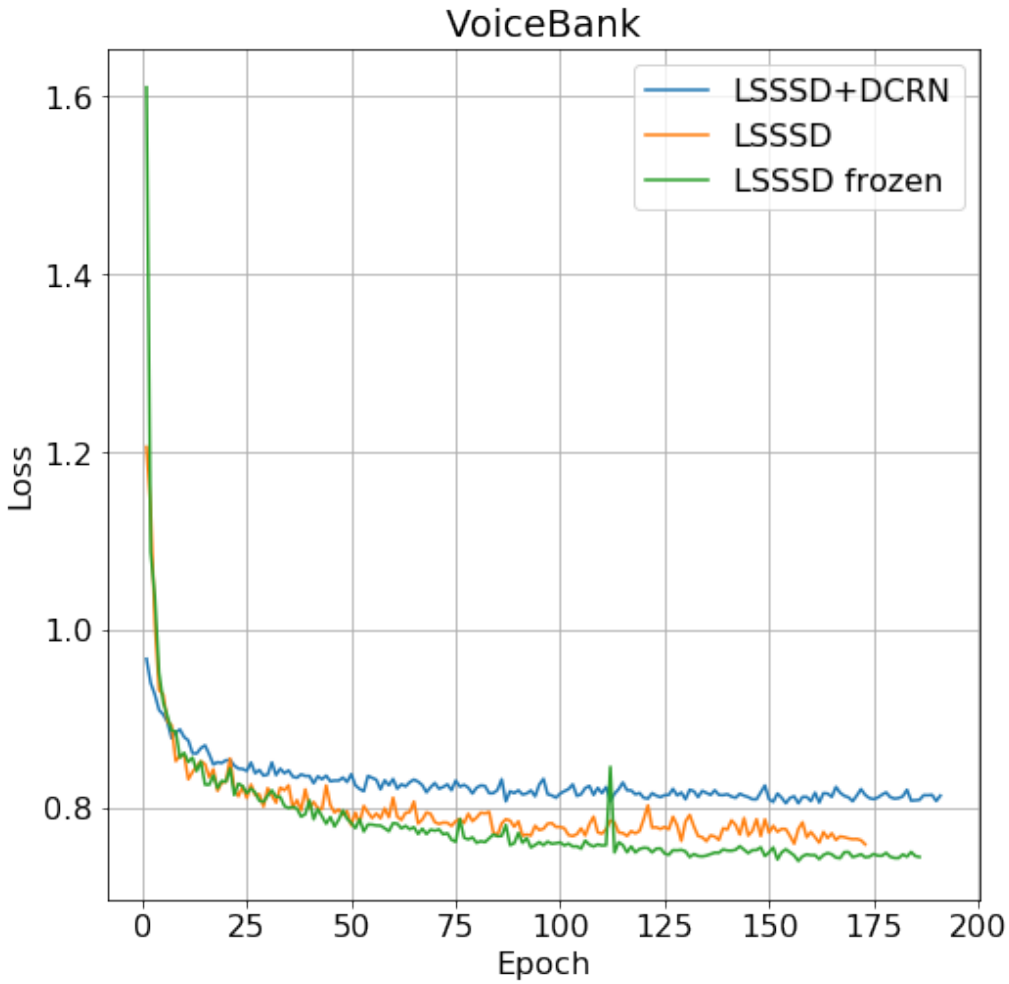

Рисунок 10 — Значения функций потерь на валидации

5.3 Результаты

По итогу проведённых экспериментов были получены модели, сравнимые по качеству как между собой, так и с базовыми моделями из статей. Более того, построенные модификации не уступают по качеству базовым моделям, а по ряду метрик даже превосходят их. На рисунках 11 и 12 представлены результаты измерений различных метрик, оценивающих качество шумоподавления для различных моделей при разных значениях SNR.

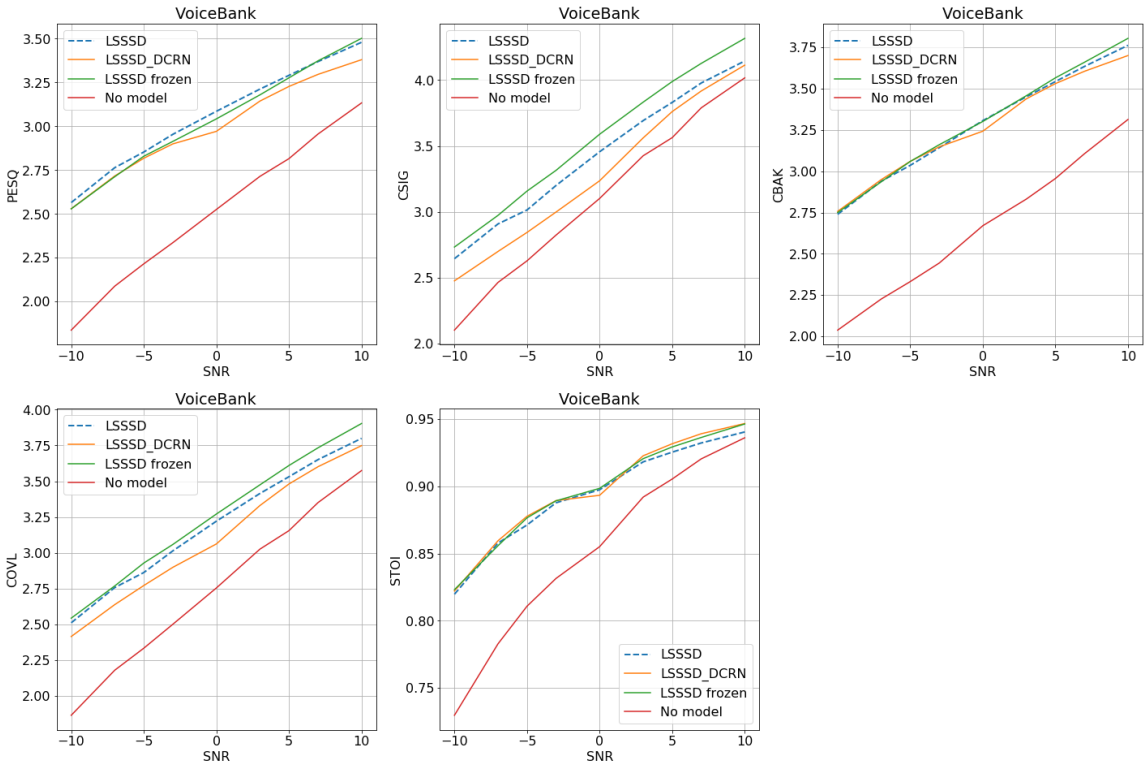

Рисунок 11 — Метрики качества шумоподавления на наборе VoiceBank

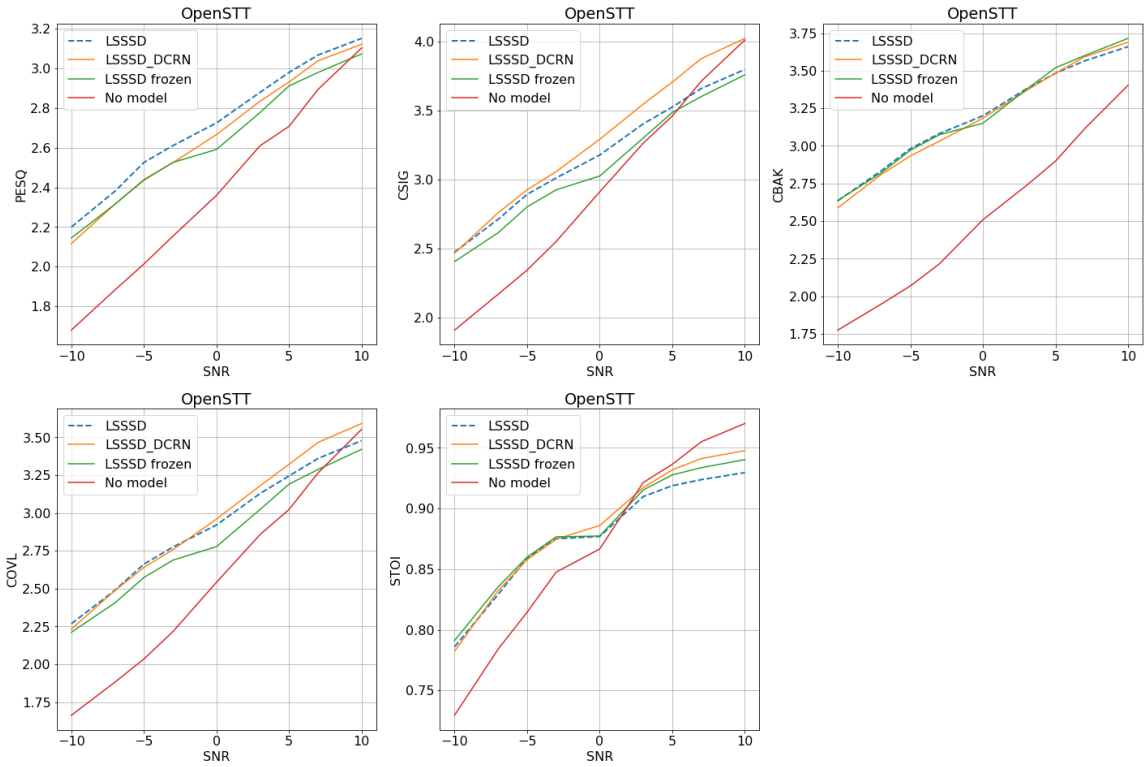

Рисунок 12 — Метрики качества шумоподавления на наборе OpenSTT+DEMAND

Более того, для всех представленных моделей шумоподавления были также проведены замеры качества на русскоязычных данных (на наборе OpenSTT). Результаты этих тестов приведены на рисунке 12. Хочется отметить сравнимость результатов на англоязычных и русскоязычных данных - в обоих случаях модели дают прирост в качестве, хотя во время обучения модели не получали данных на иных языках, помимо английского, что может говорить об инвариантности построенных моделей относительно языка и предполагать возможность их эффективного использования для речи на большинстве существующих языков. В таблицах 9 и 10 представлены усредненные по SNR результаты.

Таблица 9 — Результаты на наборе VoiceBank

	PESQ	STOI	COVL	CBAK	CSIG
LSSSD	**3,064**	0,894	3,196	3,284	3,430
LSSSD frozen	3,040	0,897	**3,255**	**3,300**	**3,559**
LSSSD + DCRN	3,000	**0,898**	3,105	3,270	3,290
No model	2,514	0,852	2,749	2,658	3,102

Таблица 10 — Результаты на наборе OpenSTT

	PESQ	STOI	COVL	CBAK	CSIG
LSSSD	**2,724**	0,879	2,925	3,202	3,183
LSSSD frozen	2,639	0,884	2,842	**3,207**	3,102
LSSSD + DCRN	2,665	**0,886**	**2,959**	3,189	**3,294**
No model	2,378	0,869	2,559	2,519	2,926

Также для всех моделей шумоподавления было измерено их влияние на качество выбранных ASR моделей. Результаты, представленные на рисунках 13 и 14 и в таблицах 11, 12 и 13 показали положительное влияние дополнительного шумоподавления на качество автоматического распознавания речи. Более того построенные модификации базовых моделей шумоподавления оказывают более сильное влияние при большей мощности наложенного шума (значения мне меньших значениях SNR)

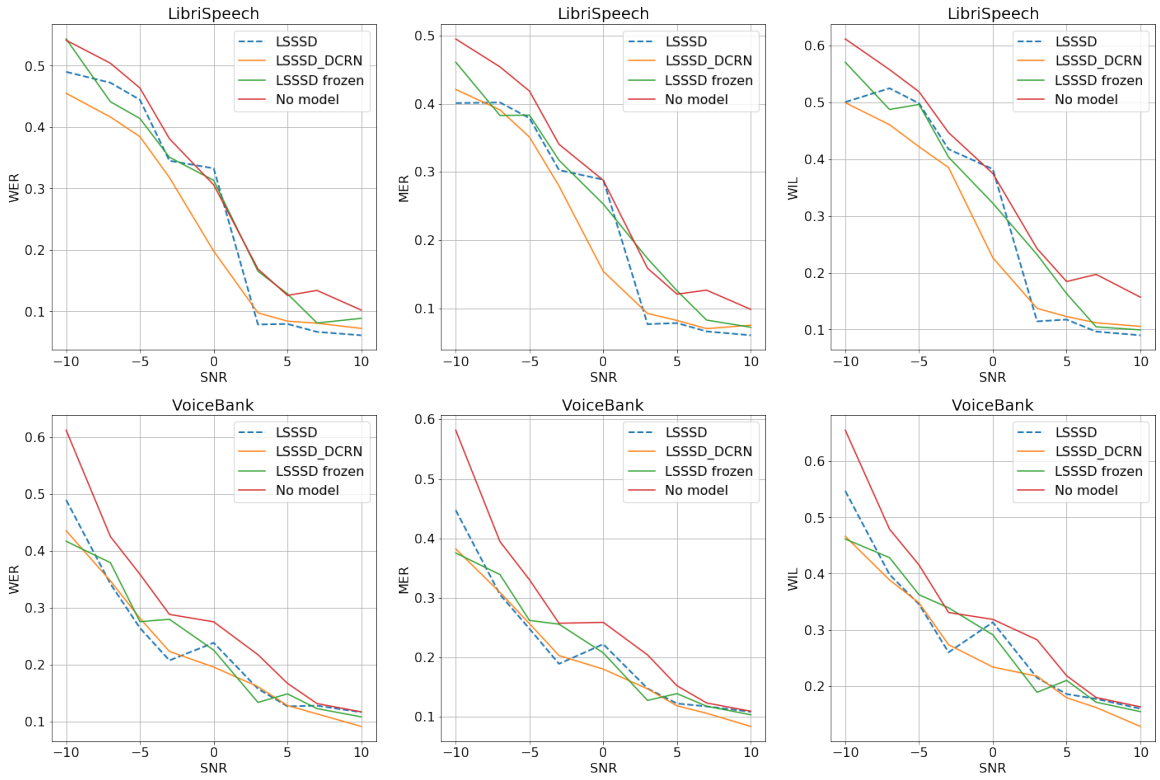

Рисунок 13 — Метрики качества распознавания речи на англоязычных наборах VoiceBank и LibriSpeech

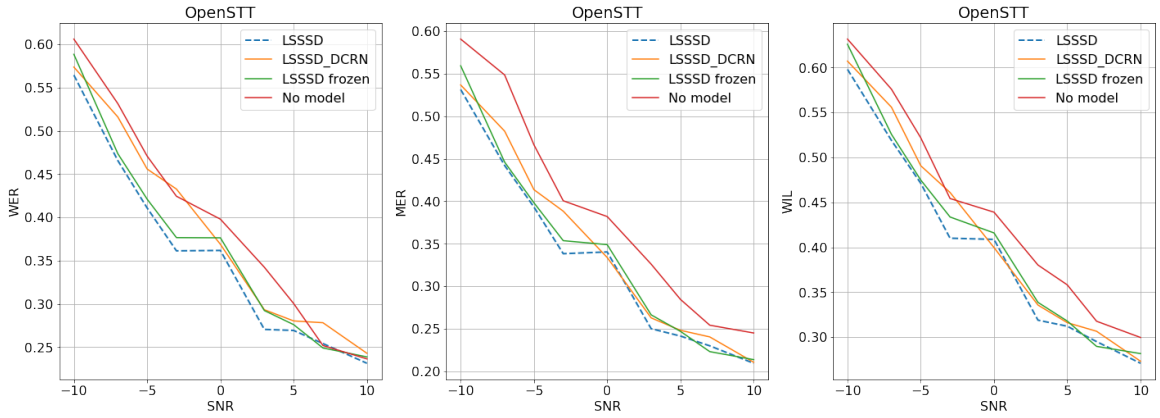

Рисунок 14 — Метрики качества распознавания речи на русскоязычном наборе OpenSTT

Таблица 11 — Результаты на наборе VoiceBank

	WER	MER	WIL
LSSSD	0,230	0,212	0,289
LSSSD frozen	0,232	0,214	0,290
LSSSD + DCRN	**0,219**	**0,198**	**0,266**
No model	0,288	0,268	0,338

Таблица 12 — Результаты на наборе LibriSpeech

	WER	MER	WIL
LSSSD	0,263	0,228	0,305
LSSSD frozen	0,281	0,250	0,320
LSSSD + DCRN	**0,234**	**0,213**	**0,275**
No model	0,303	0,278	0,366

Таблица 13 — Результаты на наборе OpenSTT

	WER	MER	WIL
LSSSD	**0,354**	**0,331**	**0,400**
LSSSD frozen	0,366	0,340	0,412
LSSSD + DCRN	0,383	0,346	0,416
No model	0,396	0,389	0,442

Стоит отметить, что положительное влияние на качество распознавания речи при использовании моделей шумоподавления оказывается на всем спектре значений SNR, даже при его больших значениях, когда можность шума крайне мала на фоне мощности сигнала с речью.

5.4 Демонстрационный стенд

В рамках подготовки данной работы был также реализован демонстрационный стенд, предназначенный для демонстрации работоспособности описанных выше алгоритмов, а также проверки всего вышесказанного на практике.

При разработке для стенда были сформулированы следующие требования:

1) обеспечивать проведение всестороннего анализа работоспособности разработанных алгоритмов в условиях произвольности входных данных;

2) обеспечить возможность изменения входных данных путем добавления произвольных шумов;

3) осуществлять настройку и реконфигурацию встроенных алгоритмов;

4) обеспечивать возможность расширения его новыми компонентами;

5) осуществлять контроль качества работы алгоритмов и моделей;

С точки зрения функционала, пользователю предлагается загрузить свою аудиозапись или выбрать одну из встроенных. Далее, также по желанию пользователя, к аудиозаписи может быть добавлен шум с желаемым уровнем мощности и/или применена модель шумоподавления. Получившиеся аудиозапись направляется на распознавание речи. Архитектура предложенного стенда изображена на рисунке 15, а на рисунке 16 - его интерфейс.

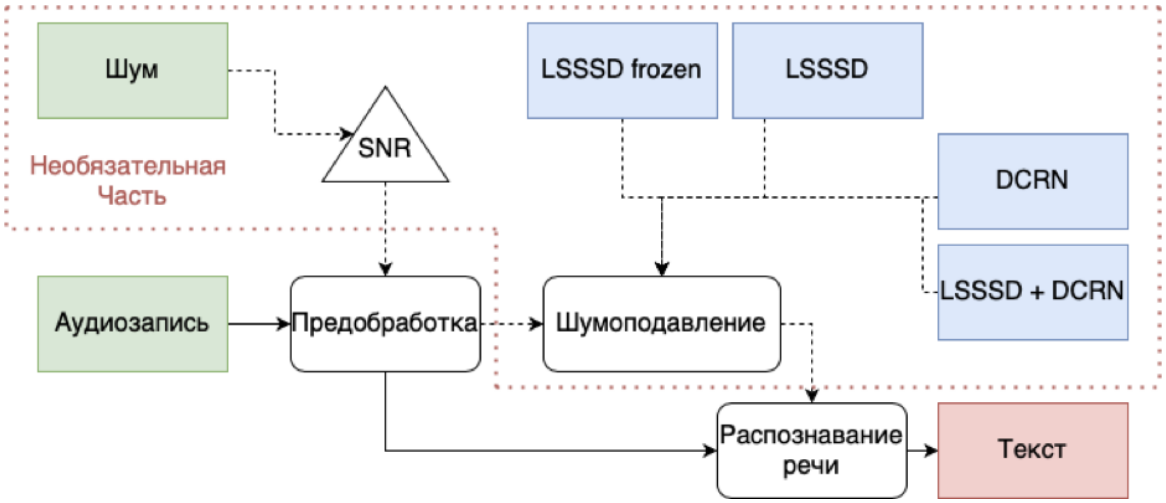

Рисунок 15 — Архитектура демонстрационного стенда

Для реализации демонстрационного стенда использовался фреймворк для веб-разработки Django - на его основе была создана серверная часть. Для интерфейса применялись языки программирования JavaScrpt и JQuery. В качестве базы данных была использована встраиваемая система СУБД SQlite.

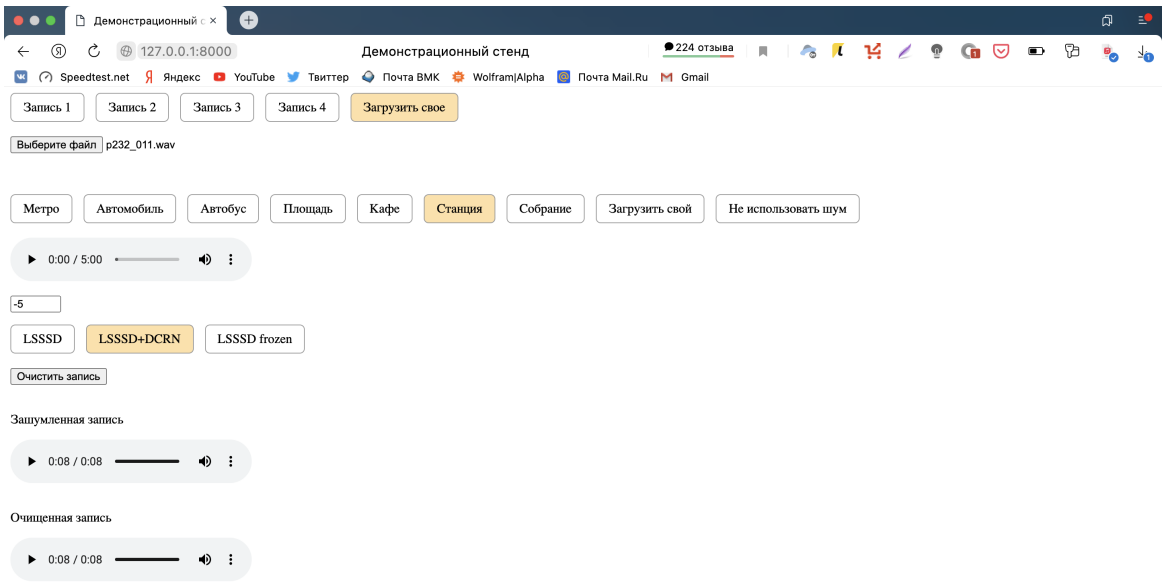

Рисунок 16 — Интерфейс демонстрационного стенда

5.5 Выводы

В практической части работы был реализован автоматизированный компонентный экспериментальный стенд. Было проведено экспериментальное сравнение предложенных в разделе 4.3 подходов к решению задачи подавления шума в аудио данных. Для каждого из рассмотренных наборов данных лучшие результаты показали следующие модели (рассматриваются значения метрики WER при значении SNR = -10):

1) Набор VoiceBank - прирост 0.2 дала модель LSSSD frozen
2) Набор LibriSpeech - прирост 0.09 дала модель LSSSD+DCRN
3) Набор VoiceBank - прирост 0.05 дала модель LSSSD

На основе результатов проведенных исследований было спроектирован и реализован демонстрационный стенд, реализующий все рассмотренные алгоритмы и позволяющий пользователю в интерактиваном режиме попробовать различные комбинации как входных данных, так и используемых алгоритмов.

6 Заключение

Был проведен обзор существующих датасетов для работы с аудиоданными, проанализированы используемые метрики качество, а также проведён обзор современных подходов по подавлению шума в аудиоданных, основанных на нейронных сетях

Были реализованы два метода шумоподавления - Listening to Sound of Silence for Speech Denoising и Dense Convolutional Recurrent Network, а также построены их модификации. Полученные результаты сравнимы с предложенными, а по ряду метрик их превосходят. Для реализации использовался язык программирования Python, а также пакеты к нему: нейросетевая библиотека pyTorch, специализированная библиотека для работы со звуком librosa.

На основе тулкита EspNet были запущены и протестированы предобученные русскоязычная и англоязычная ASR модели. Результаты тестирования подтвердили факт положительного влияния дополнительного шумоподавления на качество распознавания речи.

Был построен демонстрационный стенд, реализующий весь необходимый функционал и позволяющий проверить различные конфигурации как входных данных, так и используемых методов.

По итогам проделанной работы был подготовлен доклад на научную конференецию "Ломоносовские чтения"в 2022 году.

Список использованных источников

1. *Center, Pew Research.* Nearly half of Americans use digital voice assistants, mostly on their smartphones / Pew Research Center. — Pew Research Center.

2. *Daniel S. Park Yu Zhang, Ye Jia Wei Han Chung-Cheng Chiu Bo Li Yonghui Wu.* Improved Noisy Student Training for Automatic Speech Recognition / Ye Jia Wei Han Chung-Cheng Chiu Bo Li Yonghui Wu Daniel S. Park, Yu Zhang, Quoc V. Le. — arXiv:2005.09629v2, 2020.

3. *Joachim Thiemann Nobutaka Ito, Emmanuel Vincent.* The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings / Emmanuel Vincent Joachim Thiemann, Nobutaka Ito. — 21st International Congress on Acoustics, Acoustical Society of America, Jun 2013, Montreal, Canada.

4. *J. F. Gemmeke D. P. W. Ellis, D. Freedman A. Jansen W. Lawrence R. C. Moore M. Plakal M. Ritter.* Audioset: A large-scale dataset of manually annotated audio events / D. Freedman A. Jansen W. Lawrence R. C. Moore M. Plakal M. Ritter J. F. Gemmeke, D. P. W. Ellis. — Proc. IEEE ICASSP 2017, New Orleans, LA, 2017.

5. *Gregor Pirker Michael Wohlmayr, Stefan Petrik Franz Pernkopf.* The Pitch-Tracking Database from Graz University of Technology / Stefan Petrik Franz Pernkopf Gregor Pirker, Michael Wohlmayr. — Graz University of Technology, 2012.

6. *Valentini-Botinhao, Cassia.* Noisy speech database for training speech enhancement algorithms and TTS models / Cassia Valentini-Botinhao. — University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017.

7. *A. Ephrat I. Mosseri, O. Lang T. Dekel K. Wilson A. Hassidim W. T. Freeman M. Rubinstein.* Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation / O. Lang T. Dekel K. Wilson A. Hassidim W. T. Freeman M. Rubinstein A. Ephrat, I. Mosseri. — ACM Transactions on Graphics. ISSN 0730-0301. doi: 10.1145/3197517.3201357, 2018.

8. *Vassil Panayotov Guoguo Chen, Daniel Povey Sanjeev Khudanpur.* LIBRISPEECH: AN ASR CORPUS BASED ON PUBLIC DOMAIN AUDIO BOOKS / Daniel Povey Sanjeev Khudanpur Vassil Panayotov, Guoguo Chen. — ICASSP 2015.

9. *Anna Slizhikova Alexander Veysov, Diliara Nurtdinova Dmitry Voronin.* Russian Open Speech To Text (STT/ASR) Dataset / Diliara Nurtdinova Dmitry Voronin Anna Slizhikova, Alexander Veysov.

10. *Yi Hu, Philipos C. Loizou.* Evaluation of Objective Measures for Speech Enhancement / Philipos C. Loizou Yi Hu. — Department of Electrical Engineering University of Texas at Dallas Richardson, TX, USA.

11. Signal-to-noise ratio. — Wikipedia.

12. *Morris, Andrew.* From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition / Andrew Morris, Viktoria Maier, Phil Green. — 2004. — 10.

13. *Santiago Pascual1 Antonio Bonafonte1, Joan Serra.* SEGAN: Speech Enhancement Generative Adversarial Network / Joan Serra Santiago Pascual1, Antonio Bonafonte1. — arXiv:1703.09452v3, 2017.

14. *Dario Rethage Jordi Pons, Xavier Serra.* A Wavenet for Speech Denoising / Xavier Serra Dario Rethage, Jordi Pons. — arXiv:1706.07162v3, 2018.

15. *Ruilin Xu1 Rundi Wu1, Yuko Ishiwaka Carl Vondrick Changxi Zheng.* Listening to Sounds of Silence for Speech Denoising / Yuko Ishiwaka Carl Vondrick Changxi Zheng Ruilin Xu1, Rundi Wu1. — arXiv:2010.12013v1, 2020.

16. *Ashutosh Pandey Chunxi Liu, Yun Wang Yatharth Saraf.* DUAL APPLICATION OF SPEECH ENHANCEMENT FOR AUTOMATIC SPEECH RECOGNITION / Yun Wang Yatharth Saraf Ashutosh Pandey, Chunxi Liu. — arXiv:2011.03840v1, 2020.

17. *Hochreiter, Sepp.* Long Short-term Memory / Sepp Hochreiter, Jürgen Schmidhuber. — Neural computation, 1997. — 12. — Vol. 9. — Pp. 1735–80.

18. *Olaf Ronneberger, Philipp Fischer.* U-Net: Convolutional Networks for Biomedical Image Segmentation / Philipp Fischer Olaf Ronneberger,

Thomas Brox. — arXiv:1505.04597v1, 2015.

19. *Wenzhe Shi Jose Caballero, Ferenc Huszar Johannes Totz Andrew P. Aitken Rob Bishop1 Daniel Rueckert Zehan Wang.* Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network / Ferenc Huszar Johannes Totz Andrew P. Aitken Rob Bishop1 Daniel Rueckert Zehan Wang Wenzhe Shi, Jose Caballero. — arXiv:1609.05158v2, 2016.

20. *Sainburg, Tim.* Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires / Tim Sainburg, Marvin Thielk, Timothy Q Gentner. — Public Library of Science, 2020. — Vol. 16. — P. e1008228.

21. *Valin, Jean-Marc.* A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement / Jean-Marc Valin. — arXiv:1709.08243v3, 2018.

22. *Shinji Watanabe Takaaki Hori, Shigeki Karita Tomoki Hayashi Jiro Nishitoba Yuya Unno Nelson Enrique Yalta Soplin Jahn Heymann Matthew Wiesner Nanxin Chen1 Adithya Renduchintala1 Tsubasa Ochiai.* ESPnet: End-to-End Speech Processing Toolkit / Shigeki Karita Tomoki Hayashi Jiro Nishitoba Yuya Unno Nelson Enrique Yalta Soplin Jahn Heymann Matthew Wiesner Nanxin Chen1 Adithya Renduchintala1 Tsubasa Ochiai Shinji Watanabe, Takaaki Hori. — arXiv:1804.00015v1, 2018.

23. Chainer: a Next-Generation Open Source Framework for Deep Learning / Seiya Tokui, Kenta Oono, Shohei Hido, Justin Clayton. — Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015.

24. PyTorch: An Imperative Style, High-Performance Deep Learning Library / Adam Paszke, Sam Gross, Francisco Massa et al.; Ed. by H. Wallach, H. Larochelle, A. Beygelzimer et al. — Curran Associates, Inc., 2019. — Pp. 8024–8035.

25. Zenodo. — https://zenodo.org.

26. *Watanabe, Shinji.* Англоязычная ASR модель / Shinji Watanabe. — Zenodo, 2020.

27. *Denisov, Pavel.* Русскоязычная ASR модель / Pavel Denisov. — Zenodo, 2021.

28. *Diederik P. Kingma, Jimmy Lei Ba.* ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION / Jimmy Lei Ba Diederik P. Kingma. — arXiv:1412.6980v9, 2017.