ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет Санкт-Петербургская школа
физико-математических и компьютерных наук

Онофрийчук Илья Степанович

РАСПОЗНАВАНИЕ РЕЧИ В ДЛИННЫХ АУДИО С ИСПОЛЬЗОВАНИЕМ ТРАНСФОРМЕРОВ

Выпускная квалификационная работа - БАКАЛАВРСКАЯ РАБОТА
по направлению подготовки *01.03.02 Прикладная математика и информатика*
образовательная программа «Прикладная математика и информатика»

Рецензент	Руководитель
руководитель группы,	д.ф. - м.н., проф.,
ООО "Яндекс.Технологии"	департамент информатики
_____	_____
Трофимов Артем Владимирович	Омельченко Александр Владимирович

Санкт-Петербург 2023

Распознавание речи – важная задача, результаты которой применяются во многих сферах. Недавно системы распознавания речи достигли нового качества, путём внедрения решений, основанных на нейронных сетях. При этом, они столкнулись с проблемой распознавания длинных аудио, так как нейронные сети обучаются лишь на коротких записях. Наиболее популярными подходами адаптации на длинные аудио стали предварительная сегментация или использование оконного распознавания. Первый подход ограничивает итоговое качество транскрибации и усложняет поддержку систем распознавания речи. Второй метод часто основывается на внешнем к моделям алгоритмам или эвристикам, что огранивает системы в добавлении новых функций. Поэтому в данной работе исследуется возможность объединения частичных гипотез непосредственно через модель распознавания речи при декодировании длинного несегментированного аудио. В работе предлагается подход по транскрибации длинных аудио с помощью энкодер декодер модели, основанной на механизме внимания. По результатам экспериментом предложенный подход достигает сравнимого качества с существующими алгоритмическими решениям. При этом он имеет более гибкую структуру.

Ключевые слова: распознавание речи, длинные аудио, трансформер.

Nowadays, speech recognition is an important task in many fields. Recently, speech recognition systems have reached a new quality, by introducing solutions based on neural networks. In doing so, they faced the problem of recognizing long audio, since neural networks are trained only on short recordings. The most popular approaches for adapting to long audio have been pre-segmentation or the use of windowed recognition. The first approach limits the final quality of recognition and complicates the support of speech recognition systems. The second method often relies on algorithms or heuristics external to the models, which limits systems to adding new features. Therefore, this paper explores the possibility of combining partial hypotheses directly through a speech recognition model when decoding long unsegmented audio. This paper proposes an approach to recognize long recordings using an attention-based encoder-decoder model. According to experimental results, the proposed approach achieves comparable quality with existing algorithmic solutions. At the same time, it has a more flexible structure.

Keywords: speech recognition, long audio, transformer.

Содержание

Введение

Распознавание речи — это задача, связанная с преобразованием звукового сигнала голоса в текстовую форму. Эта задача имеет широкий спектр применений и оказывает значимое влияние на нашу повседневную жизнь. Например, распознавание речи имеет важное значение для обеспечения доступности информации, для предоставления удобного голосового интерфейса к различным системам, для анализа и обработки речевой информации. Сущностями, решающими данную задачу, являются системы распознавания речи.

В зависимости от области применения, к системам распознавания речи могут предъявляться различные требования. Данная работа опирается на требования к таким системам, которые исходят от продукта Yandex SpeechKIt:

- универсальность решения по языкам и доменам
- возможность обработки аудио произвольной длины
- возможность расширения на смежные с распознаванием речи задачам
- генерация выравнивания текста на аудио вместе с распознаванием

Системы распознавания речи могут быть построены на различных решениях, однако в последнее время произошёл переход к нейронным сетям [14]. Выведя системы распознавания речи на новый уровень, нейронные сети привнесли в данную область проблему, связанную с распознаванием длинных аудио. Причиной этому является то, что нейронные сети могут быть обучены лишь на коротких аудио. Возникает вопрос, каким образом адаптировать нейронные сети, лучшие в случае коротких аудио, на случай длинных записей.

Наиболее стандартными подходами здесь являются использование предварительной сегментации [21] или оконное распознавание аудио с последующей склейкой частичных гипотез [2, 3]. Проблема первого набора подходов заключается в том, что они значительно усложняют поддержку системы распознавания речи, так как требуют отдельную модель для сегментации, качество которой напрямую влияет на итоговое распознавание. Второй набор подходов часто основывается на внешних к модели алгоритмам или эвристикам, что ограничивает системы на добавление новых функций.

В связи с этим, актуальным является исследование подходов, которые бы позволяли бы расширять системы распознавания речи на дополнительные задачи в случае длинных аудио, а также использовали свойства самих нейронных сетей для распознавания длинных записей.

Цель и задачи

Целью данной работы является разработка подхода по декодированию длинных аудио, удовлетворяющего требованиям Yandex SpeechKit и использующим саму модель распознания речи для склейки частичных гипотез.

Среди основных задач данной работы можно выделить следующие:

- исследовать литературу о существующих подходах распознавания речи в коротких и длинных аудио
- выбрать и обучить модель для исследования
- предложить подход по декодированию длинных аудио с использованием выбранной модели
- исследовать свойства предложенного подхода в сравнении
  - с алгоритмическим решением Yandex SpeechKit
  - с публично доступными подходами из литературы

Достигнутые результаты

В рамках данной работы была выбрана модель, удовлетворяющая требованиям Yandex SpeechKit: универсальность и точное выравнивание. На бенчмарке LibriSpeech модель достигла современного качества 6.9 WER, сравнимого с лучшими транфмормерными решениями.

Для выбранной модели был предложен подход по декодированию длинных аудио, склеивающий частичные транскрипты через непосредственно декодер модели. Подход продемонстрировал сравнимое качество с алгоритмическим подходом Yandex SpeechKit и публичными решениями из фреймворка NeMo на LibriSpeech и TED LIUM 3 датасетах.

Предложенный подход был расширен на использование дополнительной текстовой информации при декодировании длинных аудио. В рамках экспериментов выявлен общий тренд по улучшению качества распознавания при данной модификации.

Структура работы

В главе 1 приводится обзор сформировавшихся подходов по распознавнию коротких аудио с помощью нейронных сетей, а также обзор существующих подходов по декодированию длинных аудио.

В главе 2 представлено описание модели, используемой для исследования, а также описан подход по декодированию с помощью неё длинных аудио в двух вариациях.

В главе 3 представлено описание рецепта обучения исследуемой модели, а также описание наборов данных для экспериментов.

В главе 4 приведены результаты проведённых экспериментов, включающих сравнение с алгоитмическим подходом из Yandex SpeechKit и публичными решениями из NeMo.

В заключении проводится анализ проделанной работы, а также возможных направлений для дальнейших исследований в этой области.

Глава 1. Обзор литературы

1.1 Модели распознования речи

Всего для распознавания речи сформировалось 3 высокоуровневых архитектуры нейронный сетей [14]: CTC [4], RNN-T [6] и AED [1].

Каждая архитектура содержит два основных компонента: аудио энкодер, который преобразует аудио в вектора признаков, и декодер, который имея выход энкодера генерирует текст. Аудио на вход энкодеру передаётся в виде мел-спектрограммы – матрицы, каждый столбец которой описывает локальные свойства части аудио. Декодер генерирует текст путём предсказания текстовых единиц – токенов, которыми могут быть символы или более сложные объекты, например, подслова [12].

Основные различия CTC, RNN-T и AED моделей выражаются в двух параметрах: авторегрессионность и явный учёт монотонного выравнивания текста на аудио в архитектуре 1.

Таблица 1: Основные архитектуры нейронных сетей по распознаванию речи

Характеристика	CTC	RNN-T	AED
Авторегрессия	-	+	+
Учёт выравнивания	В архитектуре	В архитектуре	Явно не учитывается

CTC модель, имея выход энкодера моделирует для каждого вектора распределение над токенами, которые произносились в соответствующей части аудио. Явно учитывая специфику задачи распознавания речи, данная архитектура позволяет генерировать точное выравнивание текста на аудио [5].

RNN-T модель, имея выход энкодера проходится по нему слева направо, последовательно генерируя распознавание: моделируя вероятность следующего токена, обусловленную на текущий выход энкодера и предыдущую историю токенов. Переход от одного выхода энкодера к следующему предсказывается декодером модели.

Декодер AED модели в авторегрессионом режиме последовательно предсказывает текст, используя выход энкодера с помощью механизма внимания. Каждый шаг декодера моделирует вероятность следующего токена, обусловленную на весь выход экнкодера и историю предыдущих токенов. Данная модель часто используется в качестве единой модели для разных доменов, языков, а также смежных с распознаванием речи задач [20, 29]. AED модель расширяется на дополнительные домены и языки, так как позволяет масштабировать словарь, используемый декодером. Возможность расширения AED модели на дополнительные задачи исходит из гибкой архитектуры данной сети: связь между результирующим текстом и аудио не обязательно должна быть монотонной.

Модели обучаеются с помощью стохастического градиентного спуска, минимизируя в рамках данного процесса:

$$-logP(\texttt{истинное распознавание} \mid \texttt{аудио})$$

Данные необходимые для обучения – пары: (истинное распознавание, аудио).

При достаточном количестве данных, на академических датасетах авторегрессионные модели достигают сравнимого качества, при этом CTC модель отстаёт немного. В продуктовых решениях выбор архитектуры зависит от конкретной специфики области применения.

Помимо описанных 3 архитектур, популярными являются гибридные решения, например, AED + CTC [10]. И, конечно, конкретные решения отличаются структурными блоками.

1.2 Структурные блоки моделей

С течением времени структурные блоки, используемые в нейронных сетях для распознавания речи менялись [14]. Первые решения были основаны на рекуррентных нейронных сетях. После появления трансформеров [27] структурные блоки сменились на более современные. При этом особое место в распознавании речи всегда уделялось свёрткам: существуют как полностью свёрточные решения [13, 18], так и их комбинации с предыдущими двумя структурными блоками. Последнее время на академических датасетах лучшие

результаты были достигнуты с использованием conformer блоков [7]. Несмотря на появление conformer-ов, трансформер блоки остались актуальными в продуктовых решениях, в связи с их лучшей эффективностью и активной оптимизацией под GPU.

1.3 Длинных аудио. Предварительная сегментация

Ограничение по памяти устройств, на которых происходит обучение приводит к тому, что во время данного процесса используются лишь аудио длинной 16-30 секунд. Поэтому для декодирования длинных аудио необходимо прибегать к дополнительным методикам.

Наиболее популярный подход по распознаванию речи в длинных аудио с помощью нейронных сетей заключается в предварительной сегментации. В рамках данного процесса аудио разбивается на короткие сегменты, которые могут быть распознаны обычным образом. Также, из аудио удаляются части без речи. Данная сегментация выполняется с помощью Voice Activity Detection (VAD) модели [21].

Основываясь на дополнительной VAD модели, итоговое качество распознавания длинных аудио сильно зависит от её свойств. Среди результатов работы [15] можно увидеть ухудшение качества распознавания при переходе от идеального сегментирования к сегментированию, полученному с помощью VAD модели.

Таким образом, при использовании данного подхода по распознаванию речи в длинных аудио необходимо развивать и поддерживать на необходимом уровне качество сразу двух моделей. Это усложняет процесс развития систем распознавания речи, так как в него добавляются множество дополнительных процессов, связанных с обучением, настройкой VAD модели, а также сбором для неё дополнительных данных.

1.4 Длинные аудио. Оконное распознавание

Другой набор подходов по распознаванию длинных аудио не требует предварительной сегментации и может быть применён к несегментированному аудио. Общая идея таких методов заключается в том, что аудио может быть

разбито на окна фиксированной длины. После распознавания каждого из них итоговый текст может быть получен с помощью некоторого алгоритма.

В одной из первых работ [2] с данным подходом по распознаванию длинных аудио был предложен метод, требующий 50% перекрытия между сегментами. При таком подходе аудио распознается дважды, но с разным наложением непересекающихся окон. После получения двух транскрипций авторы выравнивают их друг с другом и разрешают конфликты между транскрипциями с помощью эвристики. Этот метод может быть применен ко всем трем основным методам распознавания речи, упомянутым ранее. Проведя исследование с рекуррентными нейронными сетями, авторы этой работы продемонстрировали, что AED сети являются наиболее чувствительными к увеличению длины аудио (относительно записей из обучения) типом моделей. Однако, предложенный ими алгоритм улучшает качество транскрипции длинных аудиозаписей и делает модели AED конкурентоспособными в случае распознавания длинных аудио.

Продолжая исследования[3], авторы обновили предыдущий метод, уменьшив требуемый процент перекрытия. В новом методе каждое окно состоит из двух частей: одна без наложения, другая с наложением. Из первой части текст попадает в окончательный транскрипт как есть, а во второй он обрабатывается с помощью определенных эвристик. Данный подход применим только к CTC и RNN-T моделям, так как требует выравнивания токенов транскрипта на выход энкодера. Согласно результатам авторов, модифицированный подход сохраняет качество их предыдущего подхода, при этом увеличия эффективность распознавания.

Аналогичные методы распознавания длинных аудиозаписей с использованием разбиения аудио на фиксированные окна можно найти в популярном фреймворке NeMo[11]. Алгоритмы в нём ограничены моделями CTC и RNN-T из-за активного использования выравнивания токенов транскрипта на выход энкодера.

Первый метод, под названием **Middle token**, заключается в том, что одно окно можно рассматривать как основной аудио сегмент плюс контекстное аудио по краям. Для получения итогового распознавания из каждого сегмента выбираются только токены из основной части.

Второй метод, под названием **LCS** (Longest Common Subsequence), заключается в том, что после независимого распознавания каждого окна текст из пересечения объединяется путем поиска в нем самой длинной общей подпоследовательности, а также путём использования дополнительных эвристик.

Ещё один подход по распознаванию длинных несегментированных записей с помощью разбиения аудио на окна реализован в Yandex SpeechKit. Далее в работе данный подход называется **алгоритмическим подходом**. Он склеивает частичные гипотезы, основываясь на эвристиках и методах динамического программирования. В рамках данного подхода выстраивается граф, вершинами которого являются слова, а рёбра задают возможный порядок между ними. После распознавания очередного окна для слов из оконной гипотезы находятся ближайшие или создаются новые вершины в графе, основываясь на самом слове и его позиции в аудио. Каждой вершине задаётся вес равный количеству распознаваний данного слова в различных окнах. Для получения итогового распознавания в графе находится путь с наибольшим весом. Данный подход применим к системам, которые могут генерировать выравнивание слов на аудио.

Таким образом, методы, основанные на разбиение длинного аудио на окна фиксированной длины, позволяют избавиться от необходимости наличия VAD модели для распознавания длинных аудио. Для получения конкурентоспособных результатов эти методы распознают аудио с окнами с наложением. Однако, в текущий момент объединение частичных гипотез часто основывается на эвристиках или внешних к моделям алгоритмам. Это приводит к тому, что усложняется развитие системы распознавания речи. В продукт становится невозможно добавлять новые функции, которые не прокидываются через алгоритм склейки частичных распознаваний.

1.5 Длинные аудио. CTC/RNN-T решения

Как упоминалось ранее CTC и RNN-T модели отличаются тем, что в их архитектуре явно учитывается монотонное выравнивание текста и аудио. Это приводит к тому, что для расширения данных моделей на длинные аудио достаточно обобщить аудио энкодер на записи произвольной длины. Суще-

ствуют различные работы на данную тему, модифицирующие структурные блоки энкодера [29, 23]. Однако, данные подходы ухудшают качество модели уже на коротких аудио, и все они ограничиваются в первую очередь CTC/RNN-T решениями, которые как было упомянуто ранее не являются настолько универсальными как AED модели.

1.6 Длинные аудио. AED решения

В литературе существуют работы использующие и AED модели для декодирования длинных аудио. Данные работы по-разному используют гибкость архитектуры данного типа моделей.

В работах [26, 9] рассматривается случай предварительно сегментированного аудио. Авторы предлагают при декодировании очередного сегмента использовать несколько предыдущих фрагментов в качестве контекста. Гибкость архитектуры AED модели позволяет сделать это достаточно просто, и в итоге авторы получают улучшение в качестве в сравнении с независимым распознаванием сегментов.

В работе [20] была предложена универсальная AED модель Whisper. В рамках исследования авторы также рассмотрели возможность распознавания с помощью неё длинных аудио, основываясь на оконном распознавании. После распознавания сегмента фиксированной длины, начало окна сдвигается на конец частичной гипотезы. При декодировании нового сегмента текст из предыдущего окна используется в качестве текстового контекста. Используемый авторами подход основывается на том, что декодер AED модели вместе с токенами распознавания выдаёт временные токены. Это и является основным недостатком данного подхода по двум причинам. Во-первых, значительно усложняется набор данных, необходимый для обучения. Во-вторых, предсказываемое таким образом выравнивание имеет большую ошибку в точности.

1.7 Выводы

- Существует 3 основные высокоуровневые архитектуры нейронных сетей для распознавания речи: CTC, RNN-T и AED. Отличаясь способом

декодирования аудио, данные классы моделей применяются в различных случаях. Другие решения могут быть построены на комбинации 3 основных с заменой структурных блоков. Трансформерные блоки являются одними из наиболее актуальных на данный момент

- Одним из способов адаптации нейронных сетей на длинные аудио является подход, основанный на предварительной сегментации. С одной стороны, после сегментации распознавание длинных аудио становится простым. С другой стороны, дополнительная VAD модель, сегментирующая аудио, вносит ограничение на итоговое качество распознавания и усложняет поддержку систем распознавания речи

- Другой набор подходов, основанный на разбиении аудио на окна фиксированной длины, избегает необходимости в дополнительной модели. Однако, существующие подходы во многом основаны на внешних по отношению к моделям алгоритмах или эвристиках, что усложняет добавление в систему распознавания речи новых функций

- Существуют сквозные решения, опирающиеся на особенности архитектур CTC, RNN-T. Однако, данные подходы ограничиваются только этими двумя моделями, которые не являются настолько универсальными, как AED

Глава 2. Метод

2.1 Модель

Исследование подходов по распознанию речи в длинных аудио требует наличие модели, к которой данные подходы применяются. Выбор модели для исследования в данной работе обуславливается требованиями продукта Yandex SpeechKit к системам распознавания речи.

Одно из основных требований данного продукта – универсальность решения по языкам, доменам, а также смежным с распознаванием речи задачам. Для удовлетворения данного требования основной моделью для исследования была выбрана AED модель. Согласно литературе, такая архитектура моделей последнее время применяется как единое по языкам и различным доменам решение задачи распознавания речи.

Проблема с обычной AED моделью заключается в том, что она не умеет генерировать точное выравнивание, что является одним из требований Yandex SpeechKit. Для удовлетворения данного требования в энкодер часть AED сети встраивается CTC голова. Текстовый выход AED декодера, может быть выравнен на выход CTC с помощью Forced Alignment [5].

Как было упомянуто ранее в обзоре литературы, подход интеграции CTC головы в энкодере AED модели является популярным. Однако, в данной работе общий обход интеграции AED и CTC изменён, с целью увеличения точности выравнивания. В связи с тем, что аудио энкодеры нейронных сетей активно уменьшают длину входа с целью увеличения эффективности, для точного выравнивания необходимо, чтобы выход энкодера был достаточно гранулярным. Также, необходимо чтобы токены, которыми оперирует CTC соответствовали данной гранулярности. В частности, в данной работе энкодер AED модели разбит на 2 части 1:

- первая часть – незначительно уменьшает длину входа. Поверх данной части и находится CTC голова, которая использует символьный словарь
- вторая часть энкодера, дополнительно уменьшая длину входа, получает векторное представление аудио, которым уже оперирует декодер AED модели

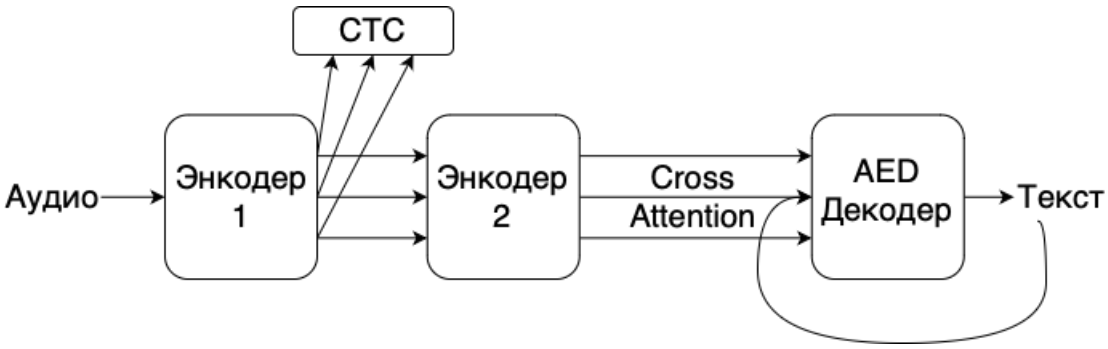

Рис. 1: Схема модели

Функция потерь, используемая для обучения AED + CTC модели, представляет собой взвешенную сумму CTCLoss [4] и кросс-ентропии.

Данные необходимые для обучения описанной модели распознаванию речи представляют собой обычные пары аудио и соответствующего текста. После обучения модель вместе с распознаванием может генерировать точное выравнивание.

2.2 Метод распознавания длинных аудио

Основная часть описанной модели (AED) имеет гибкую архитектуру, что может быть использовано при распознавании длинных аудио. В частности, в обзоре литературы упоминалась статья [9], в которой авторы предложили идею, как можно использовать гибкую архитектуру AED модели при распознавании уже сегментированного аудио. Отталкиваясь от данной идеи, но рассматривая случай несегментированных записей, данная работа предлагает собственный подход по декодированию длинных аудио с использованием AED + CTC модели.

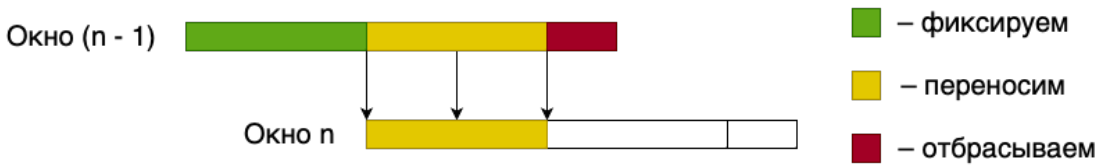

Рис. 2: Схема метода

В рамках предлагаемого подхода аудио последовательно распознаётся с помощью окон фиксированной длины. Процесс перехода от одного окна

к другому представлен на рисунке 2. В первую очередь, получив выходы энкодера для окна $n-1$, AED декодер модели генерирует для него полное распознавание, для которого также создаётся выравнивание с помощью CTC.

Текст, который попал в конец окна (сегмент, отмеченный на рисунке 2 красным), скорей всего был распознан неправильно, так как конец окна мог попасть на середину слова. Текст, попавший в данный сегмент, отбрасывается. Оставшийся текст фиксируется. Окно сдвигается на фиксированное количество секунд, но с сохранением наложения с предыдущим окном. Выходы энкодера пересчитываются на новом окне. Текст, который не был отброшен и попал в наложение, становится префиксом для декодера AED модели, при распознавании окна n. Обуславливание декодера на фиксированный префикс при распознавании очередного окна возможно за счёт гибкости архитектуры AED модели.

Гиперпараметрами подхода являются: длина окна, размер наложения и размер аудио сегмента, текст из которого отбрасывается. Предложенный подход, пользуясь свойствами AED модели, склеивает частичные распознавания непосредственно через декодер используемой сети, что является его главной особенностью.

2.3 Добавление текстового контекста

Также, предложенный подход по распознаванию речи обладает следующим свойством: при транскрибации очередного сегмента распознавание всего предыдущего аудио уже зафиксировано. Данное свойство открывает возможность использовать предшествующий текст, как дополнительную информацию при декодировании очередного окна. Предполагается, что, являясь по сути аудио языковой моделью, AED декодер таким образом обучится моделировать общий контекст, выражающийся текстом.

Интеграции в AED модели дополнительного текстового контекста предлагается достичь по аналогии с Whisper [20], путём выражения данной модификации через токены декодера. Модель обучается оперировать примерами, состоящими из двух частей, разделённых специальными токенами:

```
<prev> предыдущий текст <bos> транскрипция аудио <eos>
```

Во время обучения кросс-энтропия и CTC loss считается только по второй части.

Для осуществления подобного расширения предлагаемого подхода необходимо модифицировать и данные для обучения нейронной сети. Теперь для каждого семпла в тренировочной выборке необходимо иметь текст, который будет являться контекстом. К счастью, данное требование к данным решается достаточно просто. Часто тренировочные выборки формируются путём нарезки длинных аудио на короткие примеры. Сохранение порядка между нарезанными сегментами, позволяет удовлетворить модифицированному требованию к данным: предыдущий контекст для примера – это текст из предшествующего ему семпла. Для случая же отсутствия данного порядка далее в данной работе будет также опробована возможность обойди необходимость данной модификации данных.

2.4 Выводы

Для исследования подходов по декодированию длинных аудио, в данной работе предлагается использовать AED + CTC модель. Для удовлетворения требований продукта Yandex SpeechKit по точному выравниванию текста на аудио в данной работе модифицируется способ интеграции CTC головы в энкодере AED сети.

К AED + CTC модели предлагается использовать подход по декодированию длинных аудио, основанный на оконном распознавании. В рамках предложенного подхода частичные гипотезы склеиваются внутри декодера модели, путём переноса текста между сегментами. Также, в данной работе рассматривается возможность расширения предложенного подхода с помощью текстового контекста.

Глава 3. Постановка экспериментов

3.1 Конфигурация модели

Предложенная в данной работе архитектура сети для экспериментов использует единую конфигурацию. В качестве основных структурных блоков модель использует трансформеры [27]. Это обуславливается требованиями оптимальности по железу производительности продукта Yandex SpeechKit. В качестве блоков для уменьшения длины последовательности используются свертки [19], а более конкретно их версия из архитектуры модели распознавания речи Jasper [13].

Первый набор сверток энкодера уменьшают длину входной последовательности в два раза. Следующий за ними блок – это трансформер энкодер с 6 слоями; выход данного блока конкатенируется с его входом. CTC голова делает предсказание основываясь на данном сконкатенированном векторном представлении. Последние две части энкодера – это блок, состоящий из сверток, которые уменьшают длину последовательности в 4 раза, и следующий за ними трансформер энкодер с 16 слоями. Выход энкодера передаётся трансформер декодеру с 6 слоями. Размерность всех трансформерных блоков равна 512. Данная конфигурация модели имеет 101.65M параметров.

На вход модель получает спектрограмму с 80 мел-спектрограмма с логарифмическими величинами, посчитанным с окном 20мс и шагом 10мс. Во время обучения используются аугментации из SpecAugment [24].

Словарь декодера AED модели состоит из 5000 BPE токенов, набор которых формируется по текстам тренировочной выборки. CTC голова использует символы.

При обучении используется кросс-энтропия со сглаживающим с коэффициентом 0.1. Сглаживание оказалось обязательным механизмом регуляризации, так как без него, на небольших датасетах модель переобучается на тренировочные данные.

Коэффициент для CTCLoss равен 0.1, а его небольшая величина обуславливается тем, что CTC голова выдаёт результат по буквам, а декодер AED модели – по подсловам.

Для обучения используется AdamW оптимизатор [17] вместе с планировщиком шага обучения. Данный планировщик состоит из линейного увеличения шага обучения до 0.001 в течении первых 10000 итераций и последующим его уменьшением по функции обратного корня.

3.2 LibriSpeech

Первым датасетом для исследования является академический набор данных LibriSpeech [16]. Он состоит из аудио, в которых люди зачитывают книги на английском языке. Данный датасет отличается точно нарезанными записями, а также хорошим качеством речи в них. LibriSpeech – это наиболее популярный бенчмарк по оценке и исследованию систем распознавания речи. Данный датасет включает:

- 960 часов аудио для обучения
- 5.4 часов аудио с наиболее чистой речью в качестве тестовой выборки Test Clean
- 5.3 часов аудио с менее чистой речью – Test Other

Ключевой метрикой для оценки качества является WER (word error rate). В частности, его Total версия, которая может быть посчитана для целой выборки:

$$Total\ \ WER = \frac{I + D + S}{N}$$

, где I, D, S – количество вставок, удалений и замен слов соответственно. Данные значения считаются для каждой пары (гиптеза, истинное распознавание) с помощью расстояние левенштейна, а дальше суммируются. N – общее количество слов во всех истинных распознаваний из выборки.

На LibriSpeech предполагается оценить качество модели на коротких записях и в дополнение к этому его изменение при переходе к методам распознавания длинных аудио.

К сожалению, данный датасет не содержит полноценных записей длиной более 30 секунд. Поэтому для оценки исследуемых методов, длинные аудио создаются путём склейки нарезанных записей, объединенных по диктору

и главе. Таким образом, из тестовых выборок получаются наборы данных, описанные в таблице 2.

Таблица 2: Смоделированные конкатенацией длинные аудио LibriSpeech

Выборка	Количество длинных записей	Мин. длина, минуты	Макс. длина, минуты	Средняя длина, минуты
Test Clean	87	0.28	8.27	3.72
Test Other	90	0.67	10.05	3.56

То, как выглядят записи, описанные в таблице 2, можно представить следующим образом: из аудио, в котором один диктор зачитал целую главу книги, вырезали определённые части, а остатки склеили, сохранив порядок.

В итоге, помимо длинных аудио для оценки соответствующего качества имеем в пару те же аудио, но в нарезанном видео, подходящим для обычного распознавания. Качество модели на коротких аудио, моделирует ситуацию, когда для распознавания длинных записей мы имеем идеальное для него сегментирование. Таким образом, значения метрик на обычной тестовой выборке являются некоторой референсной оценкой для качества распознавания длинных аудио, которая может быть как улучшена, так и ухудшена.

На LibriSpeech датасете алгоритмический и предложенный подходы в первую очередь сравнивались применительно к предложенной модели. Также для выяснения состояния исследуемых подходов относительно публично доступных решений были рассмотрены комбинации подходов, представленные в таблице 3.

Для моделей из NeMo используются реализованные в этом фреймворке подходы Middle token и LCS, описанные в обзоре литературы ранее. К модели Conformer AED + CTC из ESPNET [8] применяются предложенный и алгоритмический подходы из данной работы. Помимо отличий в структурных блоках Conformer AED + CTC модель отличается от предложенной в данной работе модели видом и способом интеграции CTC в AED энкодере. Все 3 модели обучены только на LibriSpeech датасете.

Таблица 3: Публичные решения на LibriSpeech

Модель	Подход по распознаванию длинных аудио
Conformer CTC, NeMo	Middle token
Conformer RNNT, NeMo	Middle token, LCS
Conformer AED + CTC, ESPNET [8]	Предложенный, алгоритмический подходы

3.3 TED LIUM 3

Так как на LibriSpeech датасете длинные аудио были лишь смоделированы, то дополнительное исследование предполагалось провести на датасете с настоящими длинными записями. С такой целью был выбран датасет состоящий из TED выступлений, а точнее его 3 версия: TED Lium Release 3 [25]. Данный датасет включает 452 часа аудио для обучения и 3.06 часов для теста. Из-за небольшого размера тренировочной выборки данный датасет был смешан с LibriSpeech. Пересечения тренировочной и тестовой выборок отстутствуют, так как датасеты содержат данные из разных доменов.

Таблица 4: Длинные аудио TED LIUM 3

Выборка	Количество длинных записей	Мин. длина, минуты	Макс. длина, минуты	Средняя длина, минуты
Test	11	5.85	27.35	15.08

Оценить на данном датасете предполагалось алгоритмический и предложенный подходы, применительно к описанной в данной работе модели. Оценка с публичными решениями на данном датасете не проводилась в связи с отсутствием в публичном доступе моделей, обученных именно на такой комбинации данных.

На Ted Lium 3 датасете также предполагалось провести исследование интеграции текстового контекста через декодер AED модели. Конфигурация

обучения модели осталась той же, что была описана ранее. Эксперименты проводились с набором данных для обучения модели с дополнительной текстовой информацией. Во-первых, менялась вероятность, с которой к примеру добавляется предыдущий текст. Во-вторых, менялся способ по получению предыдущего текста для записи.

Первый способ получения предыдущего текста опирался на разметку, которая предоставляется с самим датасетом. В частности, для TED Lium 3 датасета для каждой записи можно определить текст, который произносился непосредственно перед ней. Для LibriSpeech можно также получить текст, который произносился перед аудио, но между полученным таким образом предыдущим текстом и текстом из записи может быть что-то ещё.

Второй способ получения предыдущего текста опирался на выравнивание, которое мы можем получить, используя CTC. В таком способе модель сначала необходимо обучить на обычных данных без текстового контекста. После этого для любой имеющейся пары: аудио, текст; можно получить выравнивание слов текста на аудио. Получив данное выравнивание, модель можно переобучить, генерируя семплы с текстовым контекстом из обычных записей. Ограничение такого подхода заключается в том, что при обучении длина текстового контекста будет ограничена длинной максимальной записи из тренировочной выборки.

Проверка последнего способа получения текстового контекста была обусловлена тем, что существуют датасеты, в которых относительный порядок между примерами не сохранён.

В работе были исследованы различные конфигурации обучения модели интеграции текстового контекста, но в следующей главе будут представлены модели со следующих конфигурациями:

- **Base**:
  - Модель, обученная стандартным образом
- **Long 0.5**:
  - Первый способ получения текстового контекста

  - Текст соответствует записи длинной 20 секунд
  - Вероятность добавления предыдущего текста: 0.5

- **Long 0.75**:
  - Первый способ получения текстового контекста
  - Текст соответствует записи длинной 20 секунд
  - Вероятность добавления предыдущего текста: 0.75

- **Short**:
  - Второй способ получения текстового контекста
  - Вероятность добавления предыдущего текста: 0.5

3.4 Синтетические данные с повторами

Помимо оценки на академических датасетах, подходы по распознаванию длинных аудио предполагалось оценить и на синтетических данных. Синтетика использовалась для того, чтобы смоделировать наиболее тяжёлый случай записей для оконных методов распознавания длинных аудио – повторения.

Для генерации данных для данного случая используется синтез. Текст для синтеза генерируется путём повторения в нём случайных чисел от 0 до 30. Для повторений используются именно числа, так как – это наиболее частый случай повторов, который, например, возникает при диктовке какого-либо номера. Кроме того, генерация не числовых повторов является более сложной задачей. Для синтеза выборки использовались 5 разных голосов. Итоговый набор данных, синтезированный для данного крайнего случая, состоит из 50 аудио, длина каждой записи равна примерно минуте.

Для оценки подходов распознавания длинных аудио на таких данных используются модели обученные на LibriSpeech. Данный подход является осмысленным, так как в тренировочной выборке датасета LibriSpeech слова из текстового представления каждого из чисел от 0 до 30 встречаются приемлемое количество раз.

3.5 Выводы

В данной главе была описана конкретная конфигурация и способ обучения исследуемой AED + CTC модели с трансформерными структурными блоками. Также, описаны конфигурации по обучению моделей интеграции текстового контекста.

Были описаны данные, на которых предполагается оценить качество подходов по декодированию длинных аудио. Во-первых, популярных бенчмарк с английской речью LibriSpeech, длинные аудио в котором моделируются путём конкатенации коротких записей. Во-вторых, TED LIUM 3 датасет, который содержит настоящие длинные TED выступления. В-третьих, синтетические данные, моделирующие наиболее тяжелый вариант аудио для оконных способов распознавания речи – повторы.

В качестве публичных подходов по декодированию длинных аудио для сравнения с предложенным методом выбраны решения из фреймворка NeMo.

Глава 4. Результаты экспериментов

В данной главе представлены результаты экспериментов, демонстрирующих качество исследуемых подходов по распознаванию длинных аудио. В рамках главы упоминаются различные варианты декодирования длинных записей, все они имеют определённый набор гиперпараметров. В течение работы над дипломом были исследованы различные комбинации данных гиперпараметров. Метрики, приведённые в данной главе, являются лучшими метриками, которых удалось достичь при вариации гиперпараметров. Данные гиперпараметры приведены в таблице 5.

Таблица 5: Гиперпараметры методов распознавния длинных аудио

Подход	Размер окна	Другие гиперпараметры
NeMo, Middle token	10 секунд	Сдвиг – 5 секунд
NeMo, LCS	10 секунд	Сдвиг – 5 секунд
Алгоритмический	10 секунд	Сдвиг – 2.5 секунды
Предложенный	10 секунд	Сдвиг – 5 секунд; Отбрасываемый текст – 1 секунда

Размер окна в 10 секунд обусловлен в том числе средней длинной аудио в тренировочных выборках исследуемых датасетов.

Декодирование коротких аудио и окон длинных аудио происходит с мощью Beam Search с размером луча 5.

4.1 LibriSpeech

Для оценки предложенного метода по распознаванию длинных аудио, в первую очередь, необходимо было оценить качество исследуемой модели на коротких аудио. Данная необходимость обуславливается тем, что качество модели на таких записях характеризует общие возможности нейронной сети по распознаванию речи. Качество модели на коротких аудио было оценено на датасете LibriSpeech, и результаты представлены в таблице 6. Все модели, представленные в таблице, обучены только на LibriSpeech датасете.

Таблица 6: Качество моделей на коротких аудио из LibriSpeech, Total WER

Модель	Параметры	Test Clean	Test Other
Conformer NeMo CTC	121M	2.5	5.7
Conformer NeMo RNN-T	120M	2.3	5.0
Conformer Espnet AED + CTC	116M	2.58	5.36
Transformer FairSeq [28] AED	268M	3.2	7.5
Transformer SpeechBrain [22] AED + CTC	72M	3.3	8.2
Proposed Transformer AED + CTC	102M	3.1	6.9

Полученные результаты на коротких аудио показывают, что в сравнении с лучшими из публично доступных трансформер решений модель достигает современных результатов. При этом, больший WER в сравнении с первыми тремя моделями обуславливается, тем что они имеют другие структурные блоки – конформеры, которые, согласно литературе, демонстрирует лучшее качество на определенных датасетах. Однако, как было упомянуто ранее, выбор трансформера в данной работе обусловлен требованиями эффективности Yandex SpeechKit.

Одним из факторов выбора описанной в данной работе модели – было необходимость в точном выравнивании. Согласно проведённым замерам, CTC из исследуемой модели выдаёт более точное выравнивание, чем CTC с токенами в виде подслов.

Убедившись в подходящем качестве исследуемой модели на коротких аудио, было оценено как изменяется качество распознавания при переходе от обычного распознавания к предложенному в данной работе методу транскрибации длинных аудио. В таблице 7 в такой постановке представлены сравнения двух подходов: предложенного и алгоритмического.

Согласно полученным результатам, оба подхода не ухудшают качество модели на нарезанных данных. При этом алгоритмический подход показывает немного лучшее качество в сравнении с предложенным подходом.

Также в таблице 7 продемонстрирована необходимость тех решений, которые были приняты в рамках разработки подхода по декодированию длинных аудио. В частности, строка "Окна без пересечения"показывает качество распо-

знавания при самом простом варианте разбиения аудио на окна. Следующая за ней строка добавляет отбрасывание текста из конца окна и улучшает качество за счёт того, что теперь меньше слов, попавших на границы, разбиваются на несколько. В итоге, качество улучшается дальше при переносе текста из одного окна в другое (строка "Предложенный метод").

Таблица 7: Качество распознавания длинных аудио, исследуемая модель, Total WER

Метод	Test Clean	Test Other
Нарезанные аудио	3.10	6.95
Алгоритмический подход	2.58	6.09
Окна без пересечения	4.26	8.42
Окна без пересечения, с отбрасыванием текста из конца окна	3.45	7.58
Предложенный метод	2.70	6.30

Для сравнения исследуемого подхода с публично доступными решениями по распознаванию длинных аудио, в аналогичной постановке были протестированы подходы из NeMo. Предложенный и алгоритмический подходы были применены к AED + CTC модели из ESPNET. Для сравнения с аналогами данная смена модели была обусловлено тем, что для полноты сравнения необходимо было уровнять исходные качества моделей на коротких аудио.

Полученные результаты 8 демонстрируют, что подходы из NeMo для оконного распознавания, незначительно ухудшают качество при переходе от распознавания нарезанных аудио. Особенно остро ситуация обстоит с RNN-T моделью и LCS подходом. Алгоритмический и предложенный подходы, применённые к ESPNET модели, не показали ухудшения качества при переходе к ним. Между двумя данными подходами сохранился тот же порядок, что был выявлен в предыдущей таблице.

Таблица 8: Качество распознавания длинных аудио, публичные решения, Total WER

Модель	Метод	Test Clean	Test Other
NeMo CTC	Нарезанные аудио	2.5	5.7
	Middle token	2.82	6.00
NeMo RNN-T	Нарезанные аудио	2.3	5.0
	NeMo RNN-T, Middle token	2.61	5.30
	NeMo RNN-T, LCS	3.29	6.44
ESPNET AED + CTC	Нарезанные аудио	2.58	5.36
	Алгоритмический подход	2.49	5.26
	Предложенный подход	2.56	5.32

Последний эксперимент, проведённый на LibriSpeech датасете был направлен на выявление устойчивости методов распознавания длинных аудио к добавлению шума. Результатом данного сравнения являются два графика. График 3 включает сравнение предложенного подхода, использующего ESPNET модель и Middle token подхода из NeMo, использующего RNN-T модель. График 4 включает сравнение предложенного и алгоритмического подходов, примененных к описанной в данной работе модели.

Для оценки подходов по декодированию длинных аудио на представленных графиках необходимо учитывать изменение качества самой модели на коротких записях при добавлении шума. На графике 3, предложенный в данной работе подход при определённом уровне шума начинает проигрывать по качеству модели на идеально нарезанных аудио. Однако, данная разность в качестве остаётся сравнимой с той разницей, которая наблюдается между двумя линиями об аналоге из NeMo. Аналогичная ситуация наблюдалась и с другими подходами оконного распознавания из данного фреймворка. Поэтому устойчивость предложенного подхода можно считать сравнимой с аналогами из NeMo.

На графике 4 предложенный подход, применённый к исследуемой модели, ведёт себя аналогично тому, что наблюдалось на предыдущем графике. Алгоритмический подход при добавлении шума сохраняет своё небольшое преимущество относительно предложенного метода. При этом обе линии остаются сравнимы друг с другом и с качеством модели на коротких аудио.

Получив общие метрики по качеству подходов, были проанализирова-

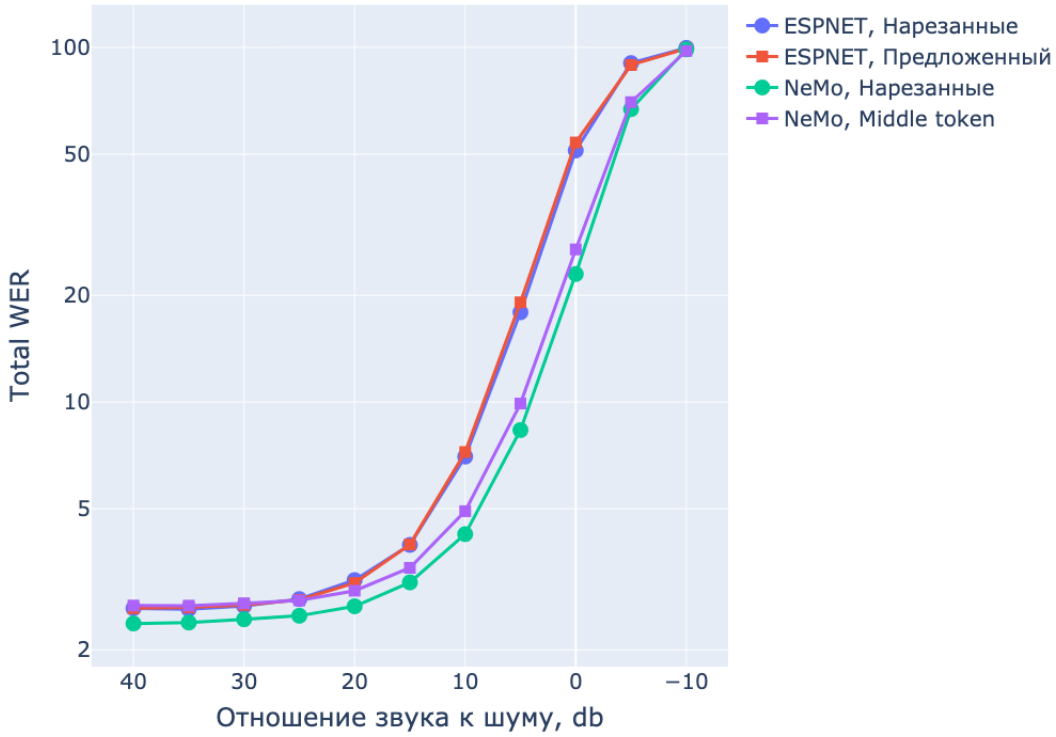

Рис. 3: LibriSpeech, Устройчивость к шуму, публичные подходы

ны конкретные примеры, где алгоритмический подход из Yandex SpeechKit оказался лучше предложенного. Было выявлено, что улучшение качества алгоритмического подхода над предложенным удаётся достичь за счёт явной агрегации гипотез из разных окон при создании итогового текста. Данное преимущество позволяет алгоритмическому подходу разменивать качество распознавания на эффективность. Предложенный же подход, являясь более эффективным, сохраняет сравнимое с алгоритмическим методом качество.

4.2 TED LIUM 3. Подтверждение результатов

Следующий набор экспериментов проводился на датасете TED Lium 3. В них использовалась только исследуемая в данной работе трансформер модель. На данном датасете ожидалось дополнительно подтвердить предыдущее сравнение алгоритмического и предложенного подходов.

Сначала необходимо было снова определить качество исследуемой моде-

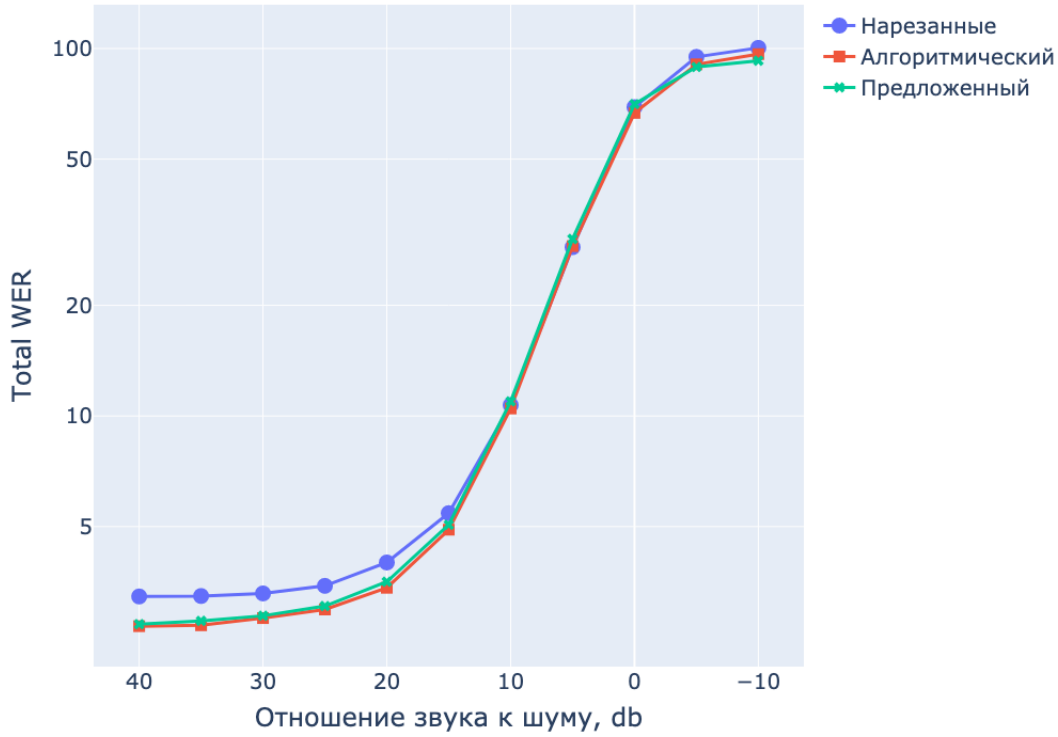

Рис. 4: LibriSpeech, Устройчивость к шуму, исследуемая модель

ли на коротких аудио 9. Модель, обученная стандартным образом, на тестовых выборках из LibriSpeech достигла, немного лучших результатов, чем модель из предыдущей части данной главы, в связи с увеличением тренировочной выборки.

Таблица 9: TED LIUM 3, Качество модели на коротких аудио. Total WER

Модель	Test Clean	Test Other	Tedlium Test
Base	2.95	6.37	5.97

После определения качества моделей на коротких аудио с Base моделью был проведён набор экспериментов, аналогичный предыдущей части данной главы.

График 5 воспроизводит результаты экспериментов с белым шумом из предыдущей части данной главы.

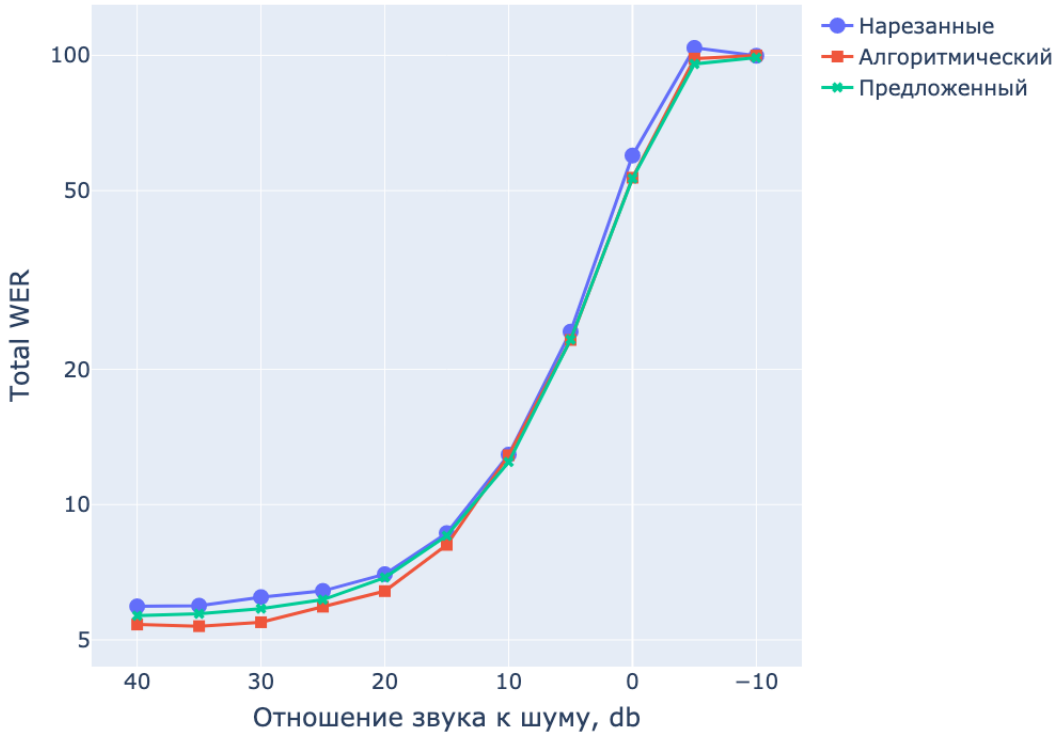

Рис. 5: Устройчивость к шуму, TED LIUM 3

4.3 Синтетические данные

В дополнение к тестовым выборкам стандартных датасетов подход был протестирован на наиболее сложном для оконных методов случае аудио – записи с повторами. Помимо повторов выборка синтетических примеров, исследуемая здесь, отличается скоростью речи дикторов 10. Предложенный подход в данном случае продемонстрировал лучшее в сравнении с алгоритмическим подходом качество, путём уменьшения процента удалений в распознавании.

Таблица 10: Синтетические данные с повторами. LibriSpeech исследуемая модель

Метод	Total WER
Алгоритмический подход	20.05
Предложенный подход	17.83

4.4 TED Lium 3. Интеграция текстового контекста

После проведения экспериментов с базовой версией предложенного алгоритма следующий блок экспериментов, описанных в данной части, был направлен на исследование интеграции текстового контекста в предложенном подходе.

Модели обучаемые вместе с дополнительным текстовым контекстом, обучились до того же качества, что и обычная модель 11.

Таблица 11: TED LIUM 3,
Модели с текстовым контекстом, короткие аудио. Total WER

Модель	Test Clean	Test Other	Tedlium Test
Base	2.91	6.29	5.87
Long 0.5	2.94	6.25	5.68
Long 0.75	2.87	6.28	5.68
Short	2.78	6.29	5.80

В таблице 12 приведены метрики качества распознавания длинных аудио из тестовой выборки TED Lium 3 датасета с использованием алгоритмического и предложенного подходов. Последние два столбца демонстрируют изменение качества предложенного метода при добавлении к нему разных длин текстового контекста: текстовая информация из предыдущих 10 или 20 секунд аудио. Таблица 13 содержит аналогичные метрики, но полученные на данных, к которым был добавлен белый шум с отношением звук к шуму 10 децибел.

Таблица 12: TED LIUM 3,
Модели с текстовым контекстом, длинные чистые аудио. Total WER

Модель	Алг. подход	Пред. метод	+ 10 сек текст	+ 20 сек текст
Base	5.34	5.59	-	-
Long 0.5	5.30	5.53	5.49	5.39
Long 0.75	5.36	5.57	5.48	5.37
Short	5.44	5.60	5.52	9.44

Таблица 13: TED LIUM 3,
Модели с текстовым контекстом, длинные шумные аудио, SNR 10 dB. Total WER

Модель	Алг. подход	Пред. метод	+ 10 сек текст	+ 20 сек текст
Base	12.87	12.43	-	-
Long 0.5	12.60	12.51	12.27	12.03
Long 0.75	12.52	12.40	12.05	11.92
Short	12.66	12.44	12.34	17.12

Для всех 4 моделей алгоритмический метод на чистых данных показал лучшие результаты, а предложенный подход сравнимое с ним качество.

Модель Short, обучаясь на записях с небольшой длинной текстовой информации, выучилась моделировать текстовый контекст, длина которого ограничена длинной текста из её обучения. Именно этим объясняется увеличение WER при добавлении в данную модель текстового контекста размера 20 секунд.

Главное вывод, который можно сделать из данных таблиц – существует общий тренд: при увеличении длины текстового контекста постепенно улучшается качество предложенного подхода. Данный тренд прослеживается как на чистых, так и на зашумленных данных. Предположительно большее изменение в качестве на зашумлённых данных объясняется тем, что при добавлении шума сигнал, идущий из аудио для генерации распознавания уменьшается, поэтому модель больше опирается на декодер.

Результаты данной части демонстрируют, что предложенный подход можно использовать вместе с текстовым контекстом. Также открывается направление для дополнительного исследования, связанного со способами интеграции текстовой информации.

4.5 Выводы

В данной главе представлены результаты экспериментов, которые продемонстрировали современное качество предложенного метода. Во-первых, было выявлено актуальное качество используемой для исследования подхода модели (6.9 WER на LibriSpeech). Во-вторых, базовый вариант исследуемого

метода показал сравнимое качество с алгоритмическим подходом и методами из NeMo на LibriSpeech датасете в двух случаях: чистых и шумных данных. Сравнимое с алгоритмическим подходом качество было подтверждено и на наборе данных TED LIUM 3. В-третьих, было выявлено, что расширение предложенного метода на интеграцию текстового контекста приводит к постепенному улучшению его качества.

Заключение

В данной работе исследовалось построение систем распознавания речи в случае длинных аудио. В современных системах распознавания речи, нейронные сети, обучаемые на коротких фрагментах речи, необходимо дополнительно адаптировать на аудио сигнал произвольной длины.

Основываясь на наработках по использованию AED моделей на длинных аудио, в данной работе был предложен собственный подход по декодированию длинных записей, главной особенностью которого является объединение частичных распознаваний непосредственно через декодер AED модели.

Предложенный метод исследовался применительно к AED + CTC модели, удовлетворяющей требованиям Yandex SpeechKit. На бенчмарке LibriSpeech модель достигает современного качества в случае коротких аудио: 6.9 WER.

По результатам экспериментом на LibriSpeech предложенный подход (6.30 WER) показал сравнимое с алгоритмическим решением (6.09 WER) качество, а также с аналогами из фреймворка NeMo. Сравнимость по качеству была продемонстрирована на чистых и шумных данных, а также на датасете TED LIUM 3.

Для дальнейшего развития предложенного подхода была рассмотрена возможность добавления в него текстового контекста. На TED LIUM 3 датасете выявлен общий тренд по улучшению итогового качества при добавлении текстовой информации: 5.57 $\rightarrow$ 5.37 на чистых данных и 12.40 $\rightarrow$ 11.92 на шумных данных.

Несмотря на то, что алгоритмические подходы могут улучшать качество путём явной агрегации множества гипотез, предложенный в данной работе подход сохраняет с ними сравнимое качество. При этом, структура предложенного подхода, его сравнимость по качеству с существующими алгоритмическими решениями, а также возможность использования текстового контекста открывает возможность к эффективному решению более сложных задач на длинных аудио, таких как:

- расстановка пунктуации
- классификация эмоций в аудио

- распознавание именованных сущностей и др.

Это и является дальнейшим направлением для исследования.

Список литературы

1. *Chan W.* Listen, Attend and Spell. — 05.08.2015. — URL: https://arxiv.org/abs/1508.01211.
2. *Chiu C.* A comparison of end-to-end models for long-form speech recognition. — 06.11.2019. — URL: https://arxiv.org/abs/1911.02242.
3. *Chiu C.* RNN-T Models Fail to Generalize to Out-of-Domain Audio: Causes and Solutions. — 07.05.2020. — URL: https://arxiv.org/abs/2005.03271.
4. Connectionist temporal classification / A. Graves [и др.] //. — 25.06.2006. — DOI: 10.1145/1143844.1143891. — URL: https://doi.org/10.1145/1143844.1143891.
5. CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition / L. Kürzinger [и др.]. — Springer Science+Business Media, 17.07.2020. — С. 267—278. — DOI: 10.1007/978-3-030-60276-5_27. — URL: https://doi.org/10.1007/978-3-030-60276-5_27.
6. *Graves A.* Sequence Transduction with Recurrent Neural Networks // arXiv (Cornell University). — 2012. — 14 нояб. — DOI: 10.48550/arxiv.1211.3711. — URL: http://arxiv.org/abs/1211.3711.
7. *Gulati A.* Conformer: Convolution-augmented Transformer for Speech Recognition. — 16.05.2020. — URL: https://arxiv.org/abs/2005.08100.
8. *Guo P.* Recent Developments on ESPnet Toolkit Boosted by Conformer. — 26.10.2020. — URL: https://arxiv.org/abs/2010.13956.
9. *Hori T.* Advanced Long-context End-to-end Speech Recognition Using Context-expanded Transformers. — 19.04.2021. — URL: https://arxiv.org/abs/2104.09426.
10. *Kim S.* Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning. — 21.09.2016. — URL: https://arxiv.org/abs/1609.06773.

11. *Kuchaiev O.* NeMo: a toolkit for building AI applications using Neural Modules. — 14.09.2019. — URL: https://arxiv.org/abs/1909.09577.

12. *Kudo T.* SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. — 19.08.2018. — URL: https://arxiv.org/abs/1808.06226.

13. *Li J.* Jasper: An End-to-End Convolutional Neural Acoustic Model. — 05.04.2019. — URL: https://arxiv.org/abs/1904.03288.

14. *Li J.* Recent Advances in End-to-End Automatic Speech Recognition. — 02.11.2021. — URL: https://arxiv.org/abs/2111.01690.

15. *Li M.* Incorporating VAD into ASR System by Multi-task Learning. — 02.03.2021. — URL: https://arxiv.org/abs/2103.01661.

16. Librispeech: An ASR corpus based on public domain audio books / V. Panayotov [и др.] //. — 19.04.2015. — DOI: 10.1109/icassp.2015.7178964. — URL: https://doi.org/10.1109/icassp.2015.7178964.

17. *Loshchilov I.* Decoupled Weight Decay Regularization. — 14.11.2017. — URL: https://arxiv.org/abs/1711.05101.

18. *Majumdar S.* Citrinet: Closing the Gap between Non-Autoregressive and Autoregressive End-to-End Models for Automatic Speech Recognition. — 05.04.2021. — URL: https://arxiv.org/abs/2104.01721.

19. *Mohamed A.* Transformers with convolutional context for ASR. — 26.04.2019. — URL: https://arxiv.org/abs/1904.11660.

20. *Radford A.* Robust Speech Recognition via Large-Scale Weak Supervision. — 06.12.2022. — URL: https://arxiv.org/abs/2212.04356.

21. *Ramírez J.*, *Górriz J. M.*, *Segura J.* Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. — 01.06.2007. — DOI: 10.5772/4740. — URL: https://doi.org/10.5772/4740.

22. *Ravanelli M.* SpeechBrain: A General-Purpose Speech Toolkit. — 08.06.2021. — URL: https://arxiv.org/abs/2106.04624.

23. *Shi Y.* Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition. — 21.10.2020. — URL: https://arxiv.org/abs/2010.10759.

24. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition / D. S. Park [и др.] //. — 18.04.2019. — DOI: 10.21437/interspeech.2019-2680. — URL: https://doi.org/10.21437/interspeech.2019-2680.

25. TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation / F. Hernandez [и др.]. — Springer Science+Business Media, 18.09.2018. — С. 198—208. — DOI: 10.1007/978-3-319-99579-3_21. — URL: https://doi.org/10.1007/978-3-319-99579-3_21.

26. Transformer-Based Long-Context End-to-End Speech Recognition / T. Hori [и др.] //. — 25.10.2020. — DOI: 10.21437/interspeech.2020-2928. — URL: https://doi.org/10.21437/interspeech.2020-2928.

27. *Vaswani A.* Attention Is All You Need. — 12.06.2017. — URL: https://arxiv.org/abs/1706.03762.

28. *Wang C.* fairseq S2T: Fast Speech-to-Text Modeling with fairseq. — 11.10.2020. — URL: https://arxiv.org/abs/2010.05171.

29. *Zhang Y.* Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. — 02.03.2023. — URL: https://arxiv.org/abs/2303.01037.