Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Соболева Дарья Михайловна

Замена живой речи на синтетическое аудио для предсказания знаков пунктуации на устройстве пользователя

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:

д.ф.-м.н.

К. В. Воронцов

Москва, 2021

Содержание

Аннотация

В данной работе представляется новая мультимодальная система прогнозирования пунктуации для английского языка, которая сочетает в себе акустические и текстовые признаки. В результате экспериментов впервые демонстрируется что, полагаясь исключительно на синтетические данные, полученные с помощью системы преобразования текста в речь с учетом интонации, можно превзойти по качеству модель, обученную на дорогостоящих аудиозаписях живого звука для решении проблемы прогнозирования пунктуации. Предложенная архитектура модели хорошо подходит для использования на устройстве. Это достигается за счет использования основанных на хэшировании эмбеддингов текстового вывода автоматического распознавания речи (ASR) в сочетании с акустическими признаками. Данные признаки далее подаются на вход квази-рекуррентной нейронной сети, что позволяет сохранять малый размер модели и низкую задержку обработки запроса.

1 Введение

Последние достижения в области автоматического распознавания речи на устройстве (Технология ASR) [1] позволяют использовать целый ряд приложений, в которых распознавание речи занимает центральное место в работе пользователя, таких как голосовая диктовка или живые субтитры [2]. Серьезным ограничением многих систем ASR является отсутствие высококачественной пунктуации.

Системы ASR обычно не предсказывают никаких знаков препинания, которые не были явно произнесены, что делает текст, транскрибированный ASR, трудным для чтения и понимания пользователями. В данной работе вводится новая система прогнозирования непроизнесенной пунктуации для уточнения выходных данных ASR. Например, имея ввод «привет Анна как дела» вместе с примерно соответствующими звуковыми сегментами, предложенная система преобразует его в «Привет, Анна! Как дела?».

Ранее было предложено несколько методов прогнозирования пунктуации. Они могут быть классифицированы на основе входных признаков: либо опираясь на акустические (интонационные) признаки [3], текстовые признаки [4, 5] или на мульти-модальном подходе, сочетающем оба типа признаков [6].

Подходы, основанные только на текстовый данных, демонстрируют низкое качества, особенно при обработке высказываний с неоднозначной пунктуацией, которые в значительной степени зависят от интонационных сигналов. И наоборот, использование аудио требует дорогостоящих аннотированных аудиозаписей живого звука. Поддержание качества данных на множестве дикторов при сборе большого количества живого звука часто является очень дорогостоящим, и представляет собой препятствие для обучения больших моделей – тем более, что знаки препинания сами по себе страдают от семантического дрейфа с течением времени [7, 8].

В данной работе предлагается объединить текстовые и акустические признаки. Чтобы уменьшить дефицит записей живого звука, предлагается использовать синтетический звук, который позволяет получать большие наборы данных, и потенциально обучать модели для областей, где в настоящее время нет записей живого человеческого звука, но существует текстовый набор данных. Учитывая недавний успех использования моделей преобразования текста в речь (Tacotron TTS) для обучения в системах ASR [9-12] в данной работе исследуются синтетические подходы к генерации звука для прогнозирования пунктуации с использованием интонационной модели Tacotron TTS [13].

Вклад данной работы заключается в следующем:

- Предложена **новая архитектура мульти-модальной акустической нейронной сети, подходящая для использования на устройстве**.
- Предложен **метод использования акустических признаков вместе с текстовыми**, значительно улучшающий качество модели при незначительном увеличении числа параметров.
- Предложен метод, позволяющий достичь качества моделей, обученных на живом человеческом аудио, для моделей, обученных **только на** 20 % **оригинальном живом человеческом аудио** (в наборе данных LibriTTS).
- Предложен **метод аугментации данных**, позволяющий полностью заменить живую речь человека на **синтезированное Tacotron TTS аудио** без потери качества для задачи предсказания пунктуации в тексте.

Насколько известно автору, это первый подход, который успешно заменяет человеческую живую речь синтетическим звуком для прогнозирования пунктуации.

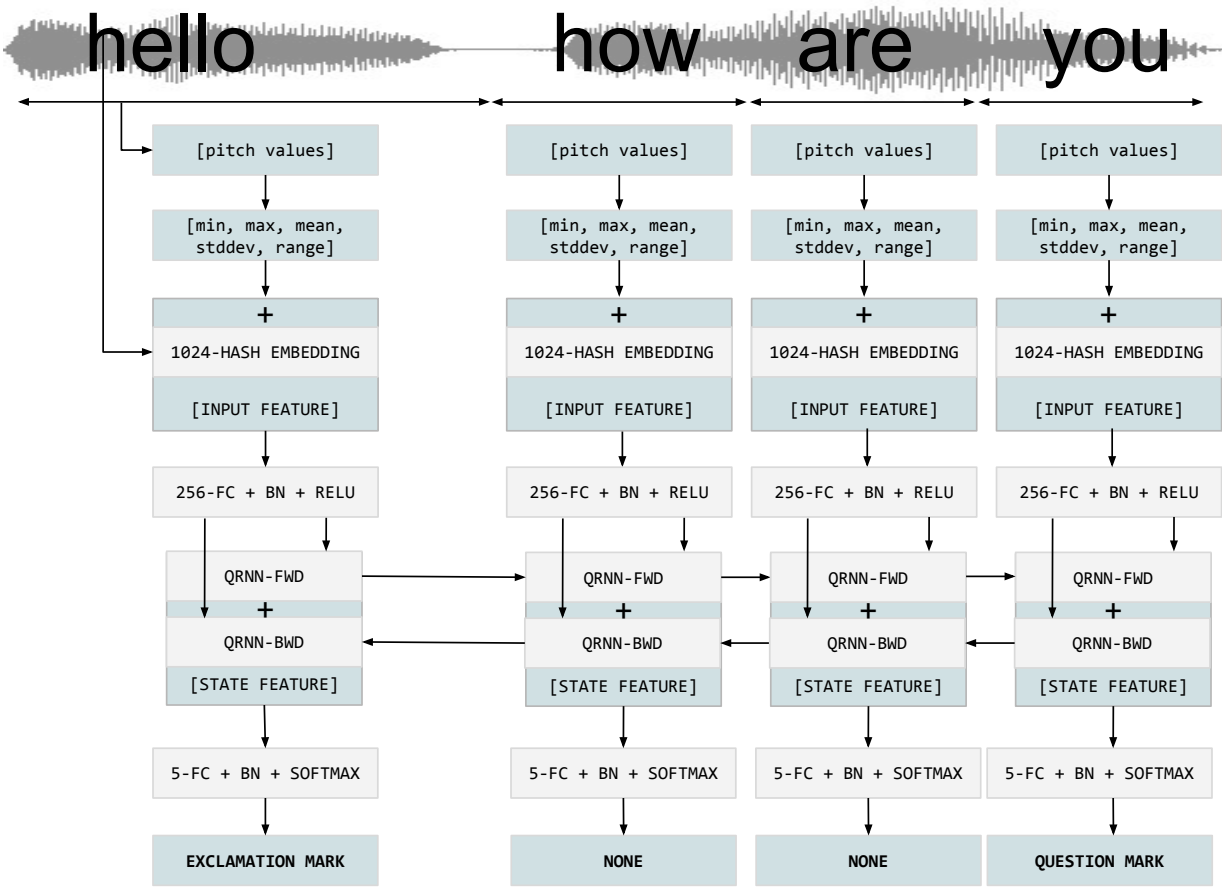

Рис. 1: Предложенная мульти-модальная архитектура. Представлена работа на примере входной последовательности. Объединяются текстовые и акустические признаки (знак конкатенации «+») и пропускаются через двунаправленный слой квази-реккурентной нейронной сети (QRNN), за которым следует полносвязный слой (FWD) с батч-нормализацией (BN) и активацией softmax. Производится классификация, какой символ пунктуации следует добавить после соответствующего входного токена.

2 Архитектура Модели

Модель использует как извлеченные значения «pitch» (высота тона) из аудио, так и выходные данные системы ASR (системы автоматического распознавания речи), обработавшей входной звук.

Высота тона или высота голоса бывает низкой или высокой. Она воспринимается ухом, и зависит от количества вибраций в секунду, производимых голосовыми связками. Высота тона является основным акустическим коррелятом тона и интонации. Высота тона или высота голоса бывает низкой или высокой. Высота тона колеблющаяся около нуля обычно является признаком паузы в речи. Необходимо заметить, что в текстовых данных пунктуация используется отчасти для разделения речи на паузы, тем самым «pitch» является релевантным признаком, описывающим исходное аудио для поставленной задачи.

Вывод ASR состоит из транскрипции высказывания, представляющей собой связный набор текста вместе с приблизительным временем произнесения каждого текстового токена внутри высказывания. Однако, временные границы между токенами только приблизительно представляют переходы от одного токена в гипотезе к следующему, и не могут использоваться для точной оценки длин пауз между ними.

Данная проблема возникает в следствие того, что текущие системы распознавания речи (например, sequence-to-sequence [1] или connectionist temporal classification [14]) не обеспечивают точного выравнивания токенов во времени, соответственно они могут быть не точными во время работы модели в режиме предсказания.

В настоящей работе предлагается использовать систему ASR, по своим характеристикам подходящую для работы на устройстве, подробное описание которой представлено в [1].

2.1 Текстовые признаки на основе хэш эмбеддингов слов

Представления для отдельных токенов определяются с помощью хэш эмбеддингов, вычисляемых с помощью операции хеширования строк, которая опирается на представление токена на уровне байтов, преобразуя его в фиксированное векторное представление данного размера [15]. В данном подходе используется детерминированная хэш функция, которая словам близким по значению присваивает хэш эмбеддинги, близкие по косинусному расстоянию.

Одной из основных мотиваций использования эмбеддингов на основе хэш функции является уменьшение размера модели. В данном подходе отсутствует необходимость хранить предобученную плотную матрицу эмбеддингов всего словаря, что является выгодным для случаев использования модели на устройстве.

Для относительно небольшого словаря размером в 20 000 токенов, даже если предположить квантованное 8-битное целочисленное представление для векторов слов, и 128-мерные векторные представления для слов, результирующая матрица увеличит размер модели примерно на 2.4 MB.

Предложенные в данной работе архитектуры используют примерно 1 MB, что является в 3 раза меньшим количеством требуемой памяти.

2.2 Оценка «pitch» для акустических признаков

Представления для звуковых сегментов, соответствующих предсказанным системой автоматического распознавания речи токенам, вычисляются с использованием оценки высоты тона («pitch») [3].

Для определения значений «pitch» из входного звукового сигнала предлагается использовать алгоритм YIN [16]. Этот метод оценивает высоту тона (частоты в Hz) для каждых 5 ms входного аудио. Предполагая частоту дискретизации 16 kHz, таким образом мы оцениваем 1 значение высоты тона для 80 аудиосэмплов. Метод оценки принимает на вход весь звуковой сигнал, соответствующий входному высказыванию.

Предсказанные оценочные значения высоты тона далее выравниваются с использованием временных границ с соответствующими входными токенами. Хотя используемая система распознавания речи не предоставляет явным образом информацию о паузах между транскрибированными токенами, метод оценки YIN выдает нулевые значения для сегментов, где звука не было зафиксировано.

Значения высоты тона, выровненные по токенам, затем используются для вычисления высокоуровневых акустических признаков для каждого текстового токена. Такая процедура необходима, так как предложенные в данной работе модели принимают на вход только признаковые описания объектов с фиксированной размерностью признакового пространства.

В качестве акустических признаков в данной работе использовались 5 скалярных статистических функций, вычисленных на высоте тона «pitch»: $mean$, $stddev$, max, min и $range$ (абсолютная разница между max и min).

Полученный вектор вычисленной акустической скалярной статистики затем объединяется с текстовыми эмбеддингами для входных токенов, и далее подается на вход моделям уже в качестве входных признаков (рис. 1).

2.3 Квази-рекуррентные слои нейронной сети

Перед передачей входных признаков в зависящий от времени слой нейронной сети предлагается спроецировать их на 256-мерный вектор, используя полносвязный батч-нормализованный слой с активацией ReLU. Последовательность проецируемых вектор-признаков затем пропускается через двунаправленные квази-рекуррентные слои нейронной сети [18]. Данные сети являются обобщением реккурентных нейронных сетей, заменяя при этом полносвязные слои на сверточные с небольшим размером ядра свертки. Такие модели являются эффективными в использовании предсказания модели на устройстве за счет использования меньшего набора параметров, и эффективности параллелизации работы.

Таблица 1: Количество объектов в обучающих, валидационных и тестовых разбиениях для набора данных LibriTTS. Для тестового разбиения также представлены частоты появления каждого из классов в задаче предсказания пунктуации: точка (PERIOD), вопросительный знак (QUESTION MARK), восклицательный знак (EXCLAMATION MARK), запятая (COMMA) и знака конца предложения (EOS), где точка (PERIOD), восклицательный знак (EXCLAMATION MARK) и вопросительный знак (QUESTION MARK) считаются взаимозаменяемыми.

Train	**Validation**	**Test**							
samples	samples	samples	tokens	punct.	EOS	PERIOD	QUESTION MARK	EXCLAMATION MARK	COMMA
42 773	2 253	1 675	24 973	3 206	1 683	1 520	97	66	1 523

Вместо dropout в данной работе применяется zoneout к скрытым узлам этого слоя [19]. При использовании zoneout предлагается не занулять случайные выходы нейронов, как это происходит при использовании dropout, вместо этого случайным выходам нейронов присваиваются значения с соответствующих выходов нейронов предыдущих слоев, что позволяет пропускать больше информации по сети.

Архитектура квази-рекуррентной нейронной сети (QRNN) работает на основе сверток. Данная модель позволяет независимо для каждого временного шага предсказывать пунктуационные символы внутри одного высказывания.

Единственный слой данной сети, который зависит от времени – это pooling-слой по каналам. В отличие от обычных слоев рекуррентные нейронных сетей (RNN), все вычисления, за исключением легковесного pooling-слоя, не зависят от предыдущих временных шагов, что позволяет быстрее делать предсказания.

Наконец, выходные данные работы двунаправленных квази-рекуррентных нейронных сетей QRNN для каждого токена объединяются и передаются для дальнейшей классификации на слой softmax. Во время обучения вычисляется кросс-энтропийную потеря с регуляризацией ℓ_2 по весам модели.

В настоящей работе решается задача многоклассовой классификации, используя выходные вероятности для 5 классов: точка (PERIOD), вопросительный знак (QUESTION MARK), восклицательный знак (EXCLAMATION MARK), запятая (COMMA) и NONE. Каждый класс, за исключением NONE, сопоставляется с соответствующим символом пунктуации, добавленным к входному токену справа. Такой подход позволяет нам выводить знаки препинания для последовательностей длиной до 100 токенов менее чем за 5 ms.

3 Методы и Эксперименты

3.1 Данные

В данной работе используется общедоступный LibriTTS датасет[1], подробное описание которого представлено в [20]. Данный корпус состоит из 585 часов английской живой человеческой речи, уменьшенной по частоте до 16 kHz. Набор данных получен из общедоступного корпуса LibriSpeech [21]. В нем сохраняется исходный текст, включая пунктуацию, но разделяется речь на границах предложений, а также удаляются высказывания со значительным фоновым шумом.

Количество объектов в обучающих, валидационных и тестовых разбиениях для набора данных LibriTTS представлено в таблице 1. Для тестового разбиения также представлены частоты появления каждого из классов в задаче предсказания пунктуации: точка (Period), вопросительный знак (Question mark), восклицательный знак (Exclamation Mark), запятая (Comma) и знака конца предложения (EoS), где точка (Period), восклицательный знак (Exclamation Mark) и вопросительный знак (Question mark) считаются взаимозаменяемыми.

Для того чтобы заменить человеческий звук синтетическим аудио, в данной работе предлагается аугментировать примеры для обучения соответствующими синтетически сгенерированными аудиосигналами. Звук синтезируется с использованием модели Tacotron, обученной на наборе данных с несколькими дикторами, а также обладающей специальной характеристикой, позволяющей синтезировать аудио с различной интонацией для примеров с одним и тем же текстовым входом [13].

Данная модель использует архитектуру кодировщика и декодировщика, получая на вход фонемные представления входной последовательности. Во время обучения модель настраивает несколько типов эмбеддингов, один из которых представляет собой вектор интонации, меняя который во время предсказания, можно получить разное произношение одного и того же набора текста. Другой важный тип эмбеддингов представляет собой вектор, описывающий диктора, который произносит текст. Таким образом модель обучается предсказывать аудио разными дикторами и с разной интонацией.

Насколько известно автору, предложенный подход аугментации данных для задачи предсказания пунктуации нигде ранее не использовался. Ранее было предложено несколько методов прогнозирования пунктуации, использующих текстовые и/или аудио признаки. Однако, системы не использующие аудио признаки демонстрировали значительно более низкое качество по сравнению с системами, обученными с использованием обоих типов признаков.

Результатов обучения на датасетах с синтезированным аудио не было представлено. Необходимо отметить, что подход с использованием живого человеческого аудио не является масштабируемым на большие наборы данных. Кроме того, существует большое разнооб-

[1] `https://www.openslr.org/60`

разие текстовых данных, открытых для использования в экспериментах. Лишь небольшой процент таких данных содержит в себе аудио записи, или какую-то информацию о произношении. Тем самым, одним из направлений экспериментов в данной работе было уменьшение процента использования живого человеческого аудио по сравнению с синтезированным без сильной потери в качестве.

3.2 Аугментация данных при помощи Tacotron TTS

Одним из естественных преимуществ синтетического аудио Text-to-Speech (Tacotron TTS) является то, что мы можем автоматически генерировать большое его количество, имея текстовый корпус и модель Tacotron TTS.

В данной работе предлагается следующий метод аугментации данных: принимая на вход текстовый набор данных, создать несколько синтетических аудиосэмплов с разной интонацией и голосами различных аудио дикторов. При данном подходе, вводятся дополнительные звуковые сигналы, улучшающие признаковое представление для обучающих данных. Данный метод позволяет увеличить размер данных для обучения в N раз, где N-максимальное количество дикторов, доступных для данной модели Tacotron TTS.

3.3 Предобработка данных

3.3.1 Аудио – Text-to-Speech

Используя модель Tacotron, предлагается генерировать звуковые сигналы, соответствующие предварительно обработанным текстовым примерам, выданным моделью ASR. Далее производится замена исходного живого человеческого звука, изначально представленного в датасете LibriTTS на соответствующую сгенерированную синтезированную версию аудио.

В данной работе предлагается провести эксперименты с 52 различными дикторами, говорящими на английском языке и являющимися его носителями, представленными в модели Tacotron TTS. Для обучения и валидации представленных моделей предлагается случайным образом разделить дикторов на 90 % для обучения и 10 % для валидационных выборок. При использовании записей живого звука из датасета LibriTTS будет сохраняться исходное разделение данных для обучения, валидации и тестирования таким образом, как это указано в [20] (таблица 1). Для тестирования моделей в данной работе используется оригинальный живой человеческий звук, представленный в датасете LibriTTS в разбиении для тестирования.

3.3.2 Предварительная обработка текста с учетом пунктуации

Необработанный текст предлагается сначала токенизировать путем разделения пробелов, знаков препинания и токенов слов, сформированнах на основе алгоритма Unicode word boundaries[2].

В данном алгоритме определение границ токенов задается в терминах упорядоченного фиксированного списка правил, указывающих состояние позиции границы (граница присутствует или отсутствует). Правила нумеруются и применяются последовательно, чтобы определить, существует ли граница с заданным смещением в заданном наборе текста. Правила обрабатываются сверху вниз. Как только правило совпадает и задает состояние позиции границы для этого смещения, процесс завершается.

Впоследствии полученные токены формируются в предложения на основании соответствующих знаков пунктуации, указанных в конце предложения (таких как, точка (Period), вопросительный знак (Question mark) и восклицательный знак (Exclamation Mark)).

Набор предложений затем представляет собой объект, который попадает в соответствующую выборку для обучения. Пример из выборки при этом может содержать несколько предложений, если общее количество токенов в нем составляет не менее 3 и не более 100, и содержит как минимум 1 знак препинания.

3.3.3 Фильтрация ошибок распознавания аудио ASR

Системы распознавания речи ASR на сегодняшний день не являются совершенными. Для того чтобы избежать смещения качества моделей предсказывающих знаки пунктуа-

[2] `https://unicode.org/reports/tr29/#Word_Boundaries`

ции в следствие ошибок ASR, предлагается выкидывать примеры из обучающего корпуса, для которых количество распознанных ASR-токенов не совпадает с количеством токенов в транскрипции, представленной в корпусе LibriTTS.

Данная мера необходима для таких объектов, так как иначе для них было бы довольно нетривиально восстановить правильную пунктуацию. Кроме того, ошибки модели предсказывающей знаки пунктуации не хотелось бы смешивать с ошибками модели распознавания речи. На этом этапе мы отфильтровываем примерно 30 % примеров обучения.

3.4 Детали обучения

3.4.1 Взвешенная Cross-Entropy

Распределение знаков препинания в тексте крайне неравномерно (Таблица 1). Чтобы смягчить данный дисбаланс, предлагается ввести взвешенную по классам кросс-энтропийную функцию потерь с весами, равными обратной частоте появления знаков препинания для каждого класса, рассчитанной на валидационной и обучающих выборках [22].

Такая модификация функции потерь гарантирует, что задачи с большим объемом доступных данных (например, такие как точка (Period) или запятая (Comma)) не будут чрезмерно доминировать в обучении. Таким образом, классы с небольшим объемом данных (например, такие как вопросительный знак (Question mark) или восклицательный знак (Exclamation Mark)) смогут быть обучены совместно с остальными, демонстрируя приемлимое качество работы на этапе тестирования.

В соответствии с PRADO [15], предлагается также добавлять ℓ_2 регуляризацию к функции потерь с весом 10^{-5}.

3.4.2 Гиперпараметры модели

Все модели используют размерность хэш эмбеддингов равную 1024, ширину ядра свертки квази-рекуррентной модели QRNN равную 7, размер скрытого состояния равный 80 и вероятность zoneout равную 0.1.

Модель обучается с использованием Adam оптимизатора [23] с начальной скоростью обучения равной 5×10^{-4}, которая экспоненциально уменьшается на 0.5 каждые 5000 шагов. Все модели обучаются в общей сложности 30 000 шагов с размером батча 512.

Модели, использующие только текстовые признаки имеют 838 127 обучаемых параметров, что увеличивается для моделей, использующих также аудио всего на 0.15 % до 839 407.

4 Результаты

4.1 Сравнение акустических и текстовых моделей

Для того чтобы мотивировать использование акустических признаков, предлагается сначала сравнить модель, использующую только текстовые признаки (0 % Human, 0 % Tacotron TTS) с акустической моделью, которая использует живой человеческий звук (100 % Human, 0 % Tacotron TTS).

Результаты, приведенные в таблице 2, демонстрируют, что акустическая модель превосходит модель, обученную только на текстовых представлениях с точки зрения точности пунктуации (punctuation accuracy) и качеству предсказанию большинства отдельных знаков препинания с точки зрения F1-score. Следует заметить, что качество предсказания запятых (Comma) становится значительно лучше. Такой результат позволяет утверждать, что модели неявно научились моделировать паузы, несмотря на отсутствие явной сегментации из предсказаний ASR.

Существует некоторое ухудшение качества предсказания с точки зрения F1-score для восклицательного знака (Exclamation Mark), который модель в основном путает с точкой (Period). Данная проблема модели будет более подробно рассмотрена в последующих экспериментах.

4.2 Звуковые комбинации живого звука и синтезированного Tacotron TTS

Учитывая обилие текстовых данных, доступных на сегодняшний день, представляется возможным обучать модели с большим количеством параметров для задач прогнозирования знаков пунктуации.

Однако из-за отсутствия разметки для живого человеческого звука мы не можем использовать акустические признаки, которые помимо улучшения качества работы всей модели также являются эффективным способом для борьбы с неоднозначностью в предсказании знаков пунктуации, когда аннотация текста пунктуацией напрямую зависит от интонации говорящего, и таким образом не может быть корректно восстановлена, полагаясь только на текстовые данные.

Для того чтобы решить эту проблему, предлагается частично заменить человеческую живую речь синтетическим звуком, синтезированным моделью Tacotron TTS, и выяснить, какого процента живого человеческого звука необходимо в обучающих данных для достижения того же качества, что и при использовании только живой человеческой речи, представленной в корпусе LibriTTS.

Для получения комбинаций с желаемым процентом для каждого типа звукового сигнала, предлагается во время предварительной обработки объектов обучения сэмплировать из соответствующего распределение Бернулли для каждого экземпляра текста. В зависимости

от значения случайной величины, принимается решение какого типа аудио необходимо использовать для аугментации данного текстового объекта.

В таблице 2 видно, что замена 100 % человеческого аудио на 100 % синтезированного аудио моделью Tacotron TTS показывает более низкое качество предсказаний запятых (Comma).

Однако, сохраняя только 20 % живого человеческого звука датасете LibriTTS, мы наблюдаем значительное улучшение качества предсказания запятых (Comma).

Эксперименты 20/80 и 50/50 демонстрируют качество на уровне 100/0 (человеческий звук). Также наблюдается улучшение F1-score для вопросительного знака (Question mark), что вероятно происходит из-за увеличения дисперсии в аудио признаках при использования дополнительных дикторов.

Таким образом, можно было бы аннотировать только 20 % корпуса LibriTTS живым человеческим звуком и использовать Tacotron TTS сгенерированное аудио для остальной части данных, сохраняя качество модели предсказания пунктуации.

Модели с 50 % использования человеческого звука демонстрируют наилучшие результаты почти для всех знаков препинания. Следует обратить внимание, что смешивание различных типов аудио приводит к неизменно более качественным моделям предсказания пунктуации, вероятно, из-за большого разнообразия данных в звуковом пространстве, например, для таких классов, как восклицательный знак (Exclamation Mark) или вопросительны знаки (Question mark). Их часто ошибочно принимают за точку (Period) из-за проблем произношения.

Использование различных дикторов увеличивает вероятность получения более разнообразных данных для обучения, что приводит к улучшению качества модели. В данной работе в экспериментах смешивание Tacotron TTS сгенерированного аудио с 40 %, 60 % и 80 % человеческого звука не дает дальнейших улучшений.

4.3 Аугментация данных при помощи Text-to-Speech аудио

Чтобы уменьшить зависимость моделей от живой человеческой речи и аннотированного аудио, предлагается полностью синтетически генерировать аудио во всем наборе данных. Для каждого обучающего примера будем выбирать случайным образом 2 из 52 Tacotron TTS-дикторов, чтобы сгенерировать два звуковых варианта одного и того же текстового примера обучения. Это позволяет получить на 200 % больший набор данных по сравнению с исходным, с большей дисперсией в пространстве аудио за счет использования нескольких дикторов.

В таблице 2 показано, что модели, обученные полностью на синтетических наборах данных с аугментацией, превосходят модели, обученные на живом человеческом звуке, в точности пунктуации (punctuation accuracy) и F1-score для восклицательного знака (Exclamation Mark).

При использовании человеческого живого звука этот класс обычно ошибочно принима-

ется за класс точки (PERIOD), и качество строго зависит от его произношения.

Используя модель Tacotron TTS, учитывающую интонацию нескольких дикторов, получаем синтетический набор данных, который лучше отражает акустические изменения, релевантные для прогнозирования пунктуации. Такие данные включают в себя большую вариативность в пространстве признаков по сравнению с исходным набором данных LibriTTS, что приводит к улучшению качества работы, обученных на нем моделей.

В проделанных экспериментах выборка из более, чем 2 дикторов не приводила к дальнейшим изменениям в качестве модели. Боле того, было замечено переобучение моделей с ростом количества используемых синтезированных аудио, аугментирующих текстовое представление объектов.

Переобучение происходило из-за того, что при одних и тех же фиксированных гиперпараметрах модели, увеличение данных в основном дублировало текстовые признаки объектов, добавляя различия только в аудио представлении. Таким образом, модель фактически была более уязвимой к запоминанию текстового корпуса для данных объектов, и все меньше демонстрировала улучшения качества на тестовых данных. Качество моделей на обучающих и валидационных данных при этом продолжало расти.

Существует решение данной проблемы, которые рассматривается как одно из направлений дальнейших исследований. Для того чтобы снизить переобучение модели, необходимо расставить веса для объектов обучающей выборки. Данные веса должны быть напрямую связаны с количеством Tacotron TTS синтезированных аудио и количеством используемых дикторов для каждого текстового экземпляра. Авторы предлагают взять веса, равные обратной частоте появления объектов с синтезированным аудио, рассчитанной по обучающей и валидационной выборке из корпуса LibriTTS.

Таблица 2: Модели оцениваются на тестовом наборе данных LibriTTS с живым человеческим звуком. В таблице представлены результаты работы моделей с точки зрения точности предсказания знаков пунктуации (Punctuation Accuracy), F1-score оценки качества отдельных знаков препинания, а также оценки качества предсказания знака конца предложения (EoS), где точка (Period), восклицательный знак (Exclamation Mark) и вопросительный знак (Question mark) считаются взаимозаменяемыми. Столбец «Human / Tacotron TTS» показывает процент объектов текста, дополненных каждым типом звука. Метрики приведены в процентах и усреднены по 3 запускам. Лучшие результаты выделены жирным шрифтом.

Human / Tacotron TTS	**Punctuation Accuracy**	**F1**				
		EoS	Period	Question mark	Exclamation Mark	Comma
0 / 0	76.35	96.90	94.42	72.62	**28.19**	51.40
100 / 0	81.56	96.87	94.09	72.87	22.20	**59.76**
50 / 50	81.67	96.94	94.38	74.75	23.52	**59.35**
20 / 80	81.38	96.76	94.21	**77.49**	22.19	**58.02**
0 / 100	81.37	96.78	93.90	73.40	18.39	57.25
0 / 200	**86.15**	96.62	94.17	76.04	**29.64**	56.22

4.4 Улучшенные акустические признаки

Для того чтобы улучшить качество наших лучших моделей из предыдущих экспериментов, обученных на 200 % Tacotron TTS синтезированном аудио (таблица 2), предлагается экспериментировать с различными типами акустических признаков.

Во-первых, предлагается рассмотреть новую технику оценки высоты тона («pitch»), которая опирается на предсказаниях предварительно обученной модели SPICE[3], полные характеристики которой описаны в [24].

Модель выводит относительные нормализованные изменения высоты тона вместе со значением неопределенности для каждого 32 ms исходного аудио (каждые 512 сэмплов для звука частотой 16 kHz). Предлагается полагаться только на нормализованные оценки изменения высоты тона при проведении экспериментов, описанных в данной работе.

Размер предварительно обученной модели SPICE составляет примерно 1.8 MB, что несомненно увеличивает размер затрачиваемой памяти на устройстве пользователя по сравнению с оценкой на основе YIN [16], используемой ранее. Что определенно является недостатком данного подхода.

Другой тип акустических признаков, который предлагается извлекать из звука представляет собой 128-мерные логарифмические энергии фильтра (признаки log-mel). Данные признаки широко используются в системах обработки речи [1, 13]. Кроме того, они представляют собой более богатое представление голосового сигнала по сравнению со значениями высоты тона («pitch»).

Таким образом, предлагается извлекать более высокоуровневые признаки, добавив еще один сверточный слой, примененный к последнему измерению признаков log-mel. Основываясь на проведенных экспериментах, сверточные слои с небольшим размером ядра, как правило, работают лучше, поэтому результаты представлены для сверточных слоев с размером ядра, равным 1.

Добавляя сверточный слой, размер модели увеличивается только примерно на 2 % до 859 983.

В таблице 3 можно видеть, что оценка высоты тона («pitch») на основе модели SPICE (Pitch-SPICE) не дает улучшения качества по сравнению с оценкой YIN (Pitch-YIN). Также утверждается, что качество моделей Pitch-SPICE потенциально может быть улучшено путем дополнительного предобучения модели SPICE на более специализированном аудио корпусе, который сильнее коррелирует с данными, используемыми в задаче предсказания пунктуации.

Однако признаки log-mel позволяют значительно сильнее повысить точность пунктуации, которые не требуют предварительного обучения дополнительных моделей. Также было замечено, что несмотря на сильный рост в качестве с точки зрения точности предсказания пунктуации (punctuation accuracy), предсказание отдельных знаков препинания стало немного хуже.

[3] `https://tfhub.dev/google/spice/1`

Таблица 3: Извлечение различных акустических признаков. Обучение на 200 % Tacotron TTS синтезированного аудио. Модели оцениваются на тестовом наборе данных LibriTTS с живым человеческим звуком. В таблице представлены результаты работы моделей с точки зрения точности предсказания знаков пунктуации (Punctuation Accuracy), F1-score оценки качества отдельных знаков препинания, а также оценки качества предсказания знака конца предложения (EoS), где точка (Period), восклицательный знак (Exclamation Mark) и вопросительный знак (Question mark) считаются взаимозаменяемыми. Метрики приведены в процентах и усреднены по 3 запускам. Лучшие результаты выделены жирным шрифтом.

Acoustic Feature	**Punctuation Accuracy**	**F1**				
		EoS	Period	Question mark	Exclamation mark	Comma
Pitch-YIN	86.15	**96.62**	**94.17**	**76.04**	**29.64**	**56.22**
Pitch-SPICE	81.54	**96.70**	**94.01**	72.60	**29.35**	55.30
Log-mel	**91.02**	95.34	92.68	63.16	24.74	41.04

Утверждается, что это происходит из-за того, что модель стремится предсказать больше знаков препинания, чем предполагается, основываясь на разметке, представленных в корпусе LibriTTS.

Во многих случаях ошибки модели были связаны с двусмысленностью предсказания пунктуации, которая заложена изначально в данных обучения. Например, предложение «привет Анна» может быть аннотировано пунктуацией по-разному, например «привет, Анна!» или «привет Анна!».

В текущей реализации неоднозначные случаи никак не отличаются от явных ошибок модели. Что существенно сказывается на качестве работы модели.

В будущей работе предлагается усовершенствовать методы интерпретации результатов работы модели для более детального определения различий между реальными ошибками и случаями неоднозначной расстановки знаков пунктуации. Также предлагается проделать большее количество экспериментов с различными представлениями признаков, основанных на акустических сигналах.

5 Заключение

Результаты, выносимые на защиту:

1. В данной работе предлагается мульти-модальная система прогнозирования пунктуации, которая использует как текстовые, так и акустические признаки, сохраняя при этом технические характеристики, позволяющие использовать ее на устройстве с небольшим размером памяти и короткой задержкой предсказания запроса.

2. Согласно проведенным экспериментам на корпусе LibriTTS, только 20 % текстовых данных должны быть аннотированы живым человеческим звуком, а остальные 80 % могут быть синтезированы с помощью модели Tacotron TTS без потери качества.

3. Опытным путем было продемонстрировано, что модели обученные на синтезированных данных с аугментацией, превосходят моделей, обученных только на живой человеческой речи. Данный результат удалось получить за счет использования Tacotron TTS модели, предсказывающей для каждого текстового примера сразу два различных аудио сигнала с двумя различными дикторами, тем самым увеличивая данные для обучения в два раза.

4. Результаты полученные в ходе проделанных экспериментов позволяют заявить, что присущее акустическим системам ограничение нехватки данных обучения для предсказания пунктуации, теперь можно ослабить за счет использования сгенерированного аудио моделью Tacotron TTS, обученной на несколько дикторов с учетом интонаций.

5. В будущем предлагается улучшать качество модель Tacotron TTS, предобучая ее на более специализированных данных (например, аудиокнигах), а также глубже исследовать различия синтезированного аудио с человеческой речью, и обучаться на больших наборах данных.

Список литературы

[1] Y. He, T. N. Sainath, R. Prabhavalkar, et al., "Streaming End-to-end Speech Recognition For Mobile Devices," 2018.

[2] M. T. Ramanovich and N. Bar, "On-Device Captioning with Live Caption," Blog post, https://ai.googleblog.com/2019/10/on-device-captioning-with-live-caption.html, October 2019.

[3] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 474–485, 2012.

[4] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: a lightweight punctuation annotation system for speech," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 1998, vol. 2, pp. 689–692 vol.2.

[5] O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," in *Interspeech*, 2016, pp. 3047–3051.

[6] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5700–5704.

[7] P. Bruthiaux, "The Rise and Fall of the Semicolon: English Punctuation Theory and English Teaching Practice," *Applied Linguistics*, vol. 16, no. 1, pp. 1–14, 03 1995.

[8] P. Bruthiaux, "Knowing when to stop: Investigating the nature of punctuation," *Language & Communication*, vol. 13, no. 1, pp. 27 – 43, 1993.

[9] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," *arXiv preprint arXiv:1905.06791*, 2019.

[10] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[11] G. Wang, A. Rosenberg, Z. Chen, Y. Zhang, B. Ramabhadran, Y. Wu, and P. Moreno, "Improving speech recognition using consistent predictions on synthesized speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7029–7033.

[12] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 301–308.

[13] RJ Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," 2018.

[14] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[15] P. Kaliamoorthi, S. Ravi, and Z. Kozareva, "PRADO: Projection attention networks for document classification on-device," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, Nov. 2019.

[16] A. D. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[17] D. Talkin and B. W. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[18] J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-Recurrent Neural Networks," 2016.

[19] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, N. R. Ke, A. Goyal, Y. Bengio, A. Courville, and C. Pal, "Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations," 2016.

[20] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. C., and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," 2019.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations (ICRL)*, Dec. 2015.

[24] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised Pitch Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.