Добрый день, я студентка группы Б17-501 Баранова Дарья. Тема моей работы – Разработка системы распознавания речи.

В процессе работы было написано около 800 строк кода.

Целью работы является разработка системы распознавания английской речи из аудиофайла (т.е черного ящика переводящего аудиофайл в текст, который был произнесен).

Речь - это наиболее естественная форма человеческого общения, поэтому технология ее распознавания является одной из ключевых для развития машинного обучения. Технология распознавания может применяться в задачах типа «человек-машина»

Исследуя предметную область, можно найти несколько популярных систем распознавания речи от крупных компаний. К их недостаткам относятся:

* Имеют закрытый исходный код
* Платное лимитированное использование
* Неизвестный алгоритм работы (самые популярные представлены на слайде, но какой именно алгоритм используется и детали реализации, оптимизация работы - неизвестны)

В этой работе для распознавания будем использовать нейросети и N-граммы.

Задача распознавания речи заключается в автоматическом определении однозначной последовательности слов, которая присутствует в исходном аудиофайле. Можно выделить несколько этапов в процессе распознавания. Каждая подзадача будет решаться несколькими способами, среди которых будет потом определяться оптимальный.

Чтобы извлечь признаки из полученного аудиосигнала:

**Спектрограмма** - диаграмма: на одной оси время, по другой оси — частота; третье измерение – амплитуда на определенной частоте в конкретный момент времени, которая представлена цветом каждой точки изображения.

Еще один способ представления это MFC. Он позволяет учитывать челевеческие особенности восприятия звука, поэтому в дальнейшем это станет оптимальным способом представления данных.

Акустическая модель состоит из рекуррентной нейронной сети. Ее архитектура будет рассмотрена позже.

Чтобы избежать задачи сегментирования акустического сигнала будем использовать алгоритм СТС. Он заключается в том, чтобы нарезать исходные данные на окна в 10 мс и подавать их на вход сети. Сеть будет выдавать матрицу вероятностей встречи определенного символа алфавита в определенный момент времени.

Пример матрицы выхода СТС представлен на слайде. По вертикали представлены символы алфавита и дополнительный пустой символ, по горизонтали – временные интервалы. Круг символизирует вероятность, что в момент времени встречается символ. Чем темнее круг, тем большую вероятность он отображает.

Для декодирования CTC используется алгоритм поиска лучшего пути, и затем операция свертки. **Таким образом, акустическая модель уже способна распознавать текст, однако она не обладает никакой информацией на уровне слов и не способна корректировать свою работу исходя из знаний о том, какие слова в языке существуют.**

Это учитывает языковая модель. Будем использовать в качестве неё нечеткий поиск по словарю. Метрикой для нечеткого поиска может выступать расстояние Левенштейна и мера жаккара.

* Расстояние левенштейна отражает различность строк.
* Мера Жаккара наоборот отражает похожесть слов. Поэтому из таблицы примеров на слайде можно видеть, что для нечеткого поиска нужно минимизировать расстояние левенштейна и максимизировать меру жаккара.

Для оценки точности распознавания существует показатели точности **(Word Error Rate) и (Character Error Rate)**. Они измеряются по одной формуле, представленной на слайде но на разном уровне – на уровне слов и на уровне символов. *Пример распознавания фразы акустической моделью, всей системой представлен на слайде.*

Разработка системы велась на Python 3 в данных средах разработки. Использовались данные библиотеки для удобства разработки.

Для определения оптимальной архитектуры нейронной сети акустической модели будем сравнивать архитектуры, представленные на графике. Третья сеть из сверточных и реккурентных слоев дала наименьшее значение потерь и продолжала обучение в течение всего времени обучения, поэтому признана оптимальной.

Как говорилось ранее, на каждом этапе распознавания нужно выделить наилучший способ решения подзадачи. На слайде представлена архитектура системы. Стрелочки соединяют именно оптимальные методы решения подзадач на каждом этапе.

Без языковой модели система дает точность распознавания WER = 40%, CER = 34%, с языковой моделью, т.е вся система целиком уже WER = 33%, CER = 29%, что является достаточно хорошим результатом. (По результатам исследований показатель WER для системы распознавания гугл составляет 18% )

Таким образом, поставленная цель работы была достигнута и была разработана система распознавания речи. В процессе работы решены следующие задачи. Достигнутая степень успешного распознавания составляет WER = 33%, CER = 29%.