Санкт–Петербургский государственный университет

Кафедра компьютерного моделирования и многопроцессорных систем

Мирошниченко Александр Сергеевич

Выпускная квалификационная работа бакалавра

Разработка системы распознавания речевых команд при помощи методов машинного обучения

Направление 01.03.02

«Прикладная математика и информатика»

Научный руководитель,

кандидат физ.-мат. наук,

доцент

Козынченко В. А.

Санкт-Петербург

2021 г.

Содержание

Введение 3

Постановка задачи 5

Обзор литературы 6

Глава 1. Теоретические сведения 7

1.1. Речь как объект распознавания 7

1.2. Речь в компьютерном представлении 8

1.3. Общая схема алгоритма распознавания 8

Глава 2. Описание решения 10

2.1. Предобработка 10

2.1.1 Нормализация сигнала 10

2.1.2 Удаление постоянной составляющей 10

2.1.3 Выделение начальной и конечной точек слова 11

2.2. Выделение речевых признаков 14

2.3. Распознавание речевых команд 14

2.3.1 Описание входных и выходных данных модели 14

2.3.2 Архитектура нейронной сети 14

2.3.3 14

2.4. Распознавание речевых команд 14

Глава 3. Результаты вычислений 15

Выводы 21

Заключение 22

Приложение 25

Введение

В современном компьютеризированном мире огромное значение имеет взаимодействие человека с компьютером - ввод и вывод информации с устройства в понятной для человека форме. Один из способов внести информацию в компьютер - записать речь через микрофон, после чего можно обрабатывать данные в памяти, которые ее представляют. Обработка может быть совершенно разной, но особенно важно распознавать слова, которые произнёс человек и давать им представление в виде текста. То есть, давать такое представление речи, как если бы она была не произнесена голосом, а напечатана при помощи клавиатуры.

Решение такой задачи может быть использовано для различных целей. К примеру, можно переписываться с другим человеком по интернету текстовыми сообщениями, при этом вообще не прикасаясь к клавиатуре, управлять различными компьютерными интерфейсами при помощи голосовых команд и т.д.

Данная проблема была актуальна со времён появления компьютеров и остаётся таковой по сей день. Особенно актуальна она стала в последнее время, когда появились качественные микрофоны, возросла мощность вычислительных устройств, увеличилось количество информации, которой обмениваются люди через интернет.

Изначально, для решения данной задачи применялись такие алгоритмы, как скрытые Марковские модели, методы динамического программирования, методы дискриминантного анализа, основанные на Байесовской дискриминации и другие. Но с появлением нейронных сетей и многочисленных экспериментов с их использованием выяснилось, что задачу распознавания речи можно решать и при помощи нейросетевого подхода. И хоть сами нейронные сети появились ещё в прошлом веке, их популярность возросла только в последнее время, в связи с ростом мощности компьютеров.

В данной работе предлагается рассмотреть решение задачи распознавания речи при помощи сверточной нейронной сети. В качестве решения подзадачи выделения характеристик речи предлагается алгоритм мел-частотных кепстральных коэффициентов, разработанный в 70-х годах прошлого века, учитывающий особенности слухового восприятия человеком.

Краткое содержание глав:

В главе 1 рассмотрены теоретические сведения о звуке, речевых признаках и общая схема алгоритма распознавания звука.

В главе 2 рассмотрен алгоритм распознавания речи. Подробно описаны подзадачи предобработки звукового сигнала, выделения речевых признаков и самого блока распознавания алгоритма.

В главе 3 приведены результаты вычислений: результаты обучения и тестирования нейронной сети, а также инструменты, которые были использованы для реализации алгоритма.

Постановка задачи

Пусть $X = \{x_1, ..., x_p, ...\}$ - множество объектов речи. Оно состоит из векторов амплитуд $x_i = \{x_i^1, ..., x_i^{k_i}\},\ i = \overline{1, \inf}$. Здесь k_i - количество записанных амплитуд в i-м объекте. У каждого объекта речи есть своя частота дискретизации w_i. Само по себе множество объектов бесконечно, однако известны значения первых p элементов. Обозначим их через $\widehat{X} = \{x_1, ..., x_p\}$.

Пусть $Y = \{y_1, ..., y_p, ...\}$ - множество скалярных меток, а $\widehat{Y} = \{y_1, ..., y_p\}$ - множество известных скалярных меток для первых p объектов речи. Таким образом известно отображение $\widehat{Z} = \widehat{X} \to \widehat{Y}$, описываемое парами значений $\widehat{Z} = \{(x_1, y_1), ..., (x_p, y_p)\}$.

Необходимо разработать алгоритм, который бы строил отображение $Z = X \to Y$.

Для того, чтобы решить поставленную задачу, необходимо разбить ее на следующие подзадачи и последовательно решить их:

- Провести предобработку первоначальных данных, содержащих звуковой сигнал в виде наборов амплитуд
- Разработать алгоритм распознавания объектов речи
- Реализовать алгоритм распознавания объектов речи
- Провести вычислительные эксперименты и выяснить, какой метод решения задачи является наиболее эффективным

В дальнейшем под объектом речи понимается отдельно взятое слово. В контексте этой работы словом является команда, которая произносится на английском языке в микрофон. Базовый набор команд составляет необходимый минимум для управления программным интерфейсом медиаплеера. В рамках данной работы сам интерфейс не рассматривается.

Обзор литературы

Первые шаги в распознавании речи были сделаны в 1952 году. Тогда трое исследователей Bell Labs - Стивен Балашек, Рулон Биддалф и Кей Дэвис - представили публике первый в истории аппарат Audrey, способный распознавать человеческую речь [1]. Это была система, позволявшая распознавать только цифры.

Позднее свои результаты в этой сфере представили компании IBM, AT&T.

Примерно в 2000 году развитие индустрии приостановилось. К этому моменту устройства были способны распознавать речь с 80% точностью. Настоящий прорыв произошел, когда компания Google представила свой голосовой поиск. Новшеством было то, что все вычисления были перенесены на мощную серверную часть, а персональные устройства отвечали только за ввод информации. С этого момента все больше компаний стали внедрять распознавание речи в свои продукты.

Среди всех работ в данной задаче стоит отметить:

1) Работа

Глава 1. Теоретические сведения

В этой главе рассмотрены теоретические сведения человеческой речи, речевых признаках и общая схема алгоритма распознавания звука.

1.1 Речь как объект распознавания

Человеческая речь - результат выдыхания человеком из себя воздуха через рот или нос. При этом воздух, проходя по трахее и бронхам, вибрирует. Саму вибрацию человек может контролировать при помощи положения языка, усиления или ослабления выдыхаемого воздушного потока и степенью натяжения различных мышц. Таким образом, речевом аппарате человека две основные составляющие - генератор тонового сигнала и совокупность фильтров.

В конечном итоге речь представляет из себя колебания воздуха - звуковые волны определенной амплитуды. Волны улавливаются мембранами в приемниках звука. В случае человеческого уха происходит следующее. Звуковые волны проходят через наружное ухо в среднее и вызывают вибрацию барабанной перепонки. Колебания с барабанной перепонки передаются на маленькие слуховые косточки в среднем ухе. А со слуховых косточек — во внутреннее ухо. Когда эти колебания достигают улитки, они воздействуют на специальные клетки — волосковые. Волосковые клетки преобразуют колебания в электрические нервные импульсы. Слуховой нерв соединяет улитку с центрами слуха в головном мозге. Когда электрические нервные импульсы достигают головного мозга, они воспринимаются как звук и обрабатываются.

В случае с компьютером и подключенным к нему микрофоном - принцип схожий. Звуковые волны колеблют мембрану в микрофоне. Разница в положении мембраны замеряется в виде электрического сигнала при помощи, например, конденсатора. Далее электрический сигнал, поступает по проводу в компьютер, и там проходит обработку.

Человеческая речь - довольно специфический звуковой сигнал. Человеческий голос в среднем имеет диапазон частот от 100Гц до 4000Гц. Чем выше частота звука, тем менее чувствительно к нему ухо человека. Еще одной особенностью является то, что для повышения громкости звука в 2 раза необходимо в 8 раз больше энергии.

Речь можно представить в виде последовательности предложений, а их в свою очередь в виде последовательности слов. Слова же состоят из фонем. В общем случае речь непрерывна, т.е. слова не отделяются друг от друга паузами, за исключением того требующих пунктационных особенностей. В этой работе не рассматривается общий случай. Здесь рассмотрен частный случай, когда речь состоит из отдельных слов, отделяемых друг от друга тишиной в понимании человеческой речи. Этот выбор был обусловлен командным типом системы распознавания, которая работает с отдельными словами.

1.2 Речь в компьютерном представлении

Каждая речевая команда с точки зрения звуковой записи в компьютере - это набор амплитудных значений, полученных с микрофона и записанных в звуковой файл.

1.3 Общая схема алгоритма распознавания

На рисунке 1 представлена общая схема работы алгоритма распознавания речевых команд. Алгоритм разделен на два основных блока, обозначенных на рисунке как Блок 1 и Блок 2. На вход алгоритму поступает наборы амплитудных значений и соответствующие частоты дискретизации в формате wav. На выходе алгоритма - текстовое представление команды.

- Блок 1 - блок, отвечающий за первоначальную обработку данных и приведение их к единому формату.

- Блок 2 - блок, отвечающий за классификацию унифицированных данных. Этот блок может работать в двух режимах: обучения и непосредственной работы. В режиме обучения меняются параметры алгоритма, которые влияют на конечный результат. В режиме непосредственной работы этого не происходит.

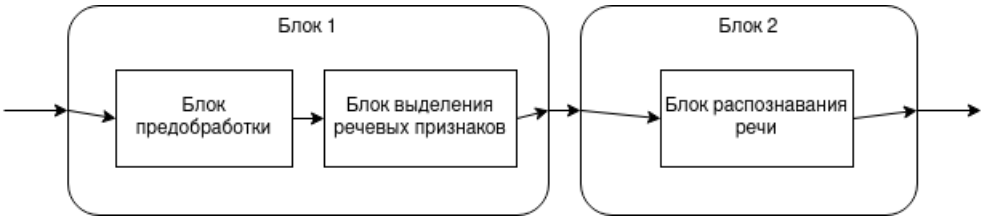

Рис. 1: Общая схема работы алгоритма распознавания речевых команд

Глава 2. Описание решения

В этой главе подробно рассмотрен весь алгоритм распознавания речевых команд. Для удобства понимая названия параграфов расположены в том же порядке, что и этапы в самом алгоритме.

2.1 Предобработка

Каждая команда - это звуковой wav файл. В каждом файле - набор амплитудных значений, которые были получены в результате записи команд дикторами.

2.1.1 Нормализация сигнала

В начале проводится нормализация амплитуд. Каждое значение амплитуды приводится к такому значению, чтобы минимум среди всех амплитуд звуковой дорожки был в 0, а максимум среди всех - в 1 по формуле:

$$\overline{x}_i = \frac{x_i}{\max_j |x_j|},\ i = \overline{1,p},\ j \in [1,p] \tag{1}$$

где x - значение амплитуды, $\overline{x}$ - новое значение амплитуды, p - количество амплитудных значений в звуковой дорожке.

Таким образом все значения амплитуд принимают значения в диапазоне $[0,1]$.

2.1.2 Удаление постоянной составляющей

Постоянная составляющая (DC-offset) - это смещение амплитуды сигнала на некоторую постоянную величину. Возникает это в аналого-цифровом сигнале из-за разницы напряжения между звуковой картой и устройством ввода. Данный эффект является помехой, от которой нужно избавиться. Для

этого необходимо вычесть из каждого значения амплитуды среднее арифметическое всех значений амплитуд по формуле:

$$\overline{x}_i = x_i - \sum_{j=1}^{p} x_j,\ i = \overline{1, p} \tag{2}$$

где x - значение амплитуды полученное на этапе нормализации, $\overline{x}$ - новое значение амплитуды, p - количество амплитудных значений в звуковой дорожке.

2.1.3 Выделение начальной и конечной точек слова

Каждая звуковая дорожка содержит в себе помимо фрагментов звукового сигнала - команды ещё и фрагменты тишины. Очень важно отделить звуковой сигнал от фрагментов тишины, т. к. именно он несёт в себе всю информацию о команде.

Для того, чтобы выделить звуковой сигнал и «обрезать» тишину в начале и в конце записи, используется алгоритм, описанный статье [6]. Каждая звуковая дорожка разбивается на фреймы - наборы амплитуд, каждый длительностью 20 мс. Начала фреймов расположены с периодичностью 10 мс. Таким образом, фреймы пересекаются между собой. Это обеспечивает целостность обработки звукового сигнала, т.е. позволяет не упустить важные фонемообразующие особенности.

Затем для каждого фрейма вычисляется мгновенная энергия:

$$E_k = \sum_{m=1}^{N} x_{k_m}^2,\ k = \overline{1, z} \tag{3}$$

где z - количество фреймов для конкретной звуковой записи, N - длина одного фрейма (количество амплитуд в одном фрейме).

Мгновенная энергия имеет один значительный недостаток. У неё очень большая чувствительность к относительно большим значениям амплитуды из-за возведения во вторую степень. Это ведёт к искажению соотношений от-

счётов звукового сигнала между друг другом. Поэтому функция мгновенной энергии переопределяется как:

$$E_k = \sum_{m=1}^{N} |x_{k_m}|,\ k = \overline{1, z} \tag{4}$$

После того, как посчитаны мгновенные энергии для каждого фрейма, вычисляется нижнее и верхнее пороговые значения:

$$\begin{aligned} I_1 &= 0.03 \cdot (MX - MN) + MN \\ I_2 &= 4 \cdot MN \\ ITL &= min(I_1,\ I_2) \\ ITU &= 10 \cdot ITL \end{aligned} \tag{5}$$

где MN, MX - минимум и максимум мгновенной энергии среди всех фреймов соответственно, ITL, ITU - нижнее и верхнее пороговое значение.

Происходит поиск фрейма, с которого начинается слово с самого первого фрейма. Фрейм, в котором значение мгновенной энергии превышает ITL, предварительно помечается как начало слова. Затем начиная с этого помеченного фрейма происходит поиск фрейма, в котором значение мгновенной энергии превышает ITU. Если значение мгновенной энергии для какого-то фрейма во время последнего поиска меньше ITL, то этот фрейм становится предварительным началом слова.

Аналогично происходит поиск конца слова в звуковой дорожке. Только поиск по фреймам происходит не с начала сигнала, а с конца.

После этого этапа имеются два предварительно помеченных фрейма m_1, m_2 - начало и конец слова в звуковом файле.

Функция мгновенной энергии, определённая формулой (4) хорошо справляется с отделением звонких звуков от тишины. Но вот глухие она отделяет плохо. Поэтому используется вторая характеристика для доопределения на-

чала и конца слова - число переходов через ноль. Это количество таких случаев, когда соседние отсчёты (значения амплитуд) имеют противоположные знаки. Определяется формулой:

$$Z_k = \frac{1}{2}\sum_{m=2}^{N} |sgn(x_{k_{m-1}}) - sgn(x_{k_m})|,\ k = \overline{1, z} \tag{6}$$

Подразумевается, что первые 100 мс звуковой записи - это тишина, и речь начинается позднее.

Вычисляется среднее значение переходов через ноль в течение первых 100 мс (7), среднее квадратическое отклонение количества переходов через ноль в течение первых 100 мс (8):

$$IZC = \frac{1}{z}\sum_{k=1}^{z} Z_k \tag{7}$$

$$\sigma_{IZC} = \sqrt{\frac{1}{z}\sum_{k=1}^{z}(Z_k - IZC)^2} \tag{8}$$

а затем пороговую функцию числа переходов через ноль по формуле:

$$IZCT = min(IF,\ IZC + 2\sigma_{IZC}), \tag{9}$$

где IF - фиксированное количество переходов через ноль (25 пересечений за 10 мс).

Происходит уточнение точек начала и конца слова в звуковой дорожке. Начиная от фрейма m_1 влево происходит поиск фреймов, у которых число переходов через ноль выше порогового значения. Поиск происходит всего на расстоянии 25 фреймов, так как производится уточнение границ слова. Если пороговое значение было превышено 3 или более раз, то фрейм r_1, где это произошло впервые, помечается как начало слова.

Аналогично от фрейма m_2 происходит поиск вправо для уточнения

точки конца слова.

На выходе этого алгоритма - 2 помеченных фрейма r_2. В итоге, сигнал обрезается, и остаётся только речевая команда в виде набора фреймов $[r_1, ..., r_2]$.

2.2 Выделение речевых признаков

Для того, чтобы выделить речевые признаки, используется алгоритм MFCC [5]

2.3 Распознавание речевых команд

2.3.1 Описание входных и выходных данных модели

2.3.2 Архитектура нейронной сети

2.3.3

2.4 Распознавание речевых команд

Глава 3. Результаты вычислений

Для решения задачи был выбран язык программирования Python 3.8. Предобработку данных было решено реализовывать при помощи Python 3.8. В качестве библиотеки для реализации нейронной сети была выбрана библиотека Keras, включенная в библиотеку Tensorflow 2.4.1.

Для создания датасета был разработан веб-сервис на NodeJS, Javascript, HTML, CSS, который позволяет записывать команды и сохранять их в нужном для программы предобработки формате. Также это позволило записать необходимое количество дикторов, которые смогли довольно быстро наговорить команды.

Было проведено 3 вычислительных эксперимента для нейронной сети типа CNN. Структура сети приведена на рисунке 2.

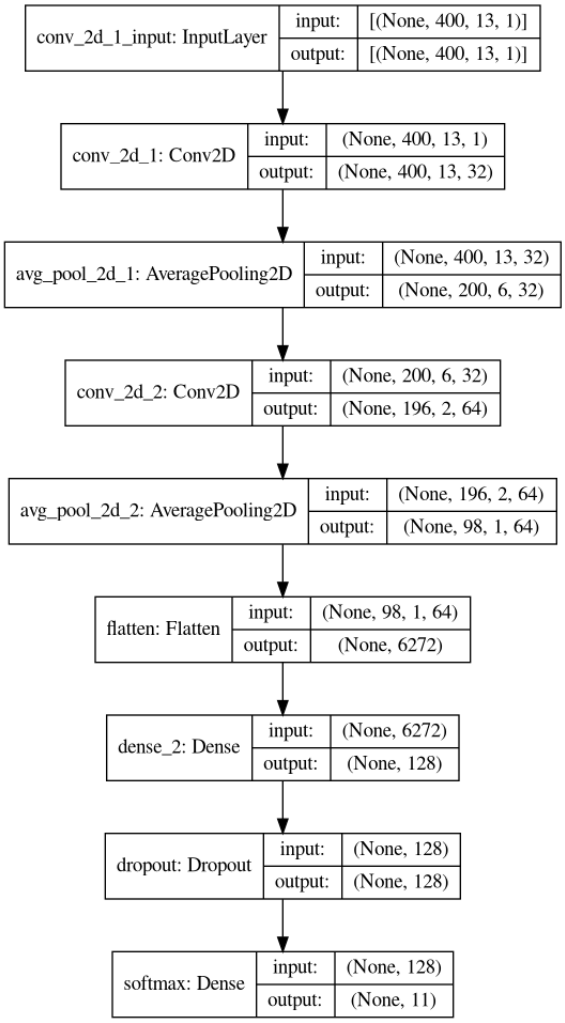

Рис. 2: Структура модели нейронной сети типа CNN

Все звуковые файлы были предобработаны при помощи алгоритма MFCC. Звуковая дорожка делится на фреймы. Каждый фрейм - отрезок звуковой дорожки длительностью 20 мс. Каждый фрейм начинается с момента (10 мс. $\times$ номер_фрейма), нумерация начинается с 0. Для каждой звуковой дорожки количество фреймов - 400. Если количество фреймов у дорожки меньше 400, то слева и справа добавляются нули. Это число было выбрано как максимально возможное количество фреймов для всех дорожек. Количество коэффициентов в алгоритме MFCC - 13, количество фильтров - 26. В итоге размерность данных, поступающих на вход нейронной сети - 400×13.

Датасет состоит из 6 дикторов. Каждый диктор записал 11 команд : 'back', 'down', 'menu', 'off', 'on', 'open', 'play', 'power', 'stop', 'up', 'volume'.

Диктор	Тип голоса	Кол-во звук. дорожек на каждую команду	Сумм. кол-во звук. дорожек
speaker1	Мужской	50	550
speaker2	Мужской	40	440
speaker3	Мужской	40	440
speaker4	Мужской	40	440
speaker5	Мужской	50	550
speaker6	Женский	50	550

Первый эксперимент: нейронная сеть обучается на первом дикторе с мужским голосом, тестирование производится на каждом дикторе.

Второй эксперимент: нейронная сеть обучается на всех дикторах с мужским голосом, тестирование производится на каждом дикторе.

Третий эксперимент: нейронная сеть обучается на всех дикторах, тестирование производится на каждом дикторе.

Датасет предварительно разделяется на тренировочную и тестовую части. На тренировочную часть отводится 70% данных диктора, на тестовую часть - 30%. В процессе тренировки после каждой эпохи тренировочные данные перемешиваются. 15% тренировочных данных в каждой эпохе - валидационные. В качестве метрики для оценки эффективности была выбрана метрика точности (accuracy), а для валидации - функция потерь категориальной кросс-энтропии (val_loss). Алгоритм оптимизации - Adam. Максимальное количество эпох - 50. Если значение метрики val_loss не уменьшается в течение 20 эпох, то обучение останавливается.

Графики обучения для каждого из экспериментов приведены на рисунках 3, 4, 5.

В конце каждого эксперимента проводится тестирование нейронной сети. А в случае обучения на all_speakers помимо тестирования производится построение матрицы ошибок (confusion matrix) для каждого диктора и для

каждого из четырех пороговых значений: 0.5, 0.6, 0.7, 0.8. Матрицы представлены на рисунке 6.

Результаты тестирования представлены в таблице 1. Обозначения, которые используются:

all_speakers = [speaker1, speaker2, speaker3, speaker4, speaker5, speaker6]

all_male_speakers = [speaker1, speaker2, speaker3, speaker4, speaker5]

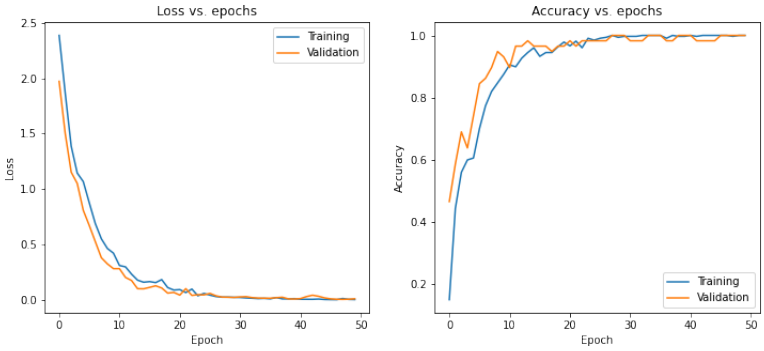

Рис. 3: Графики функции потерь и точности в течение обучения на speaker1

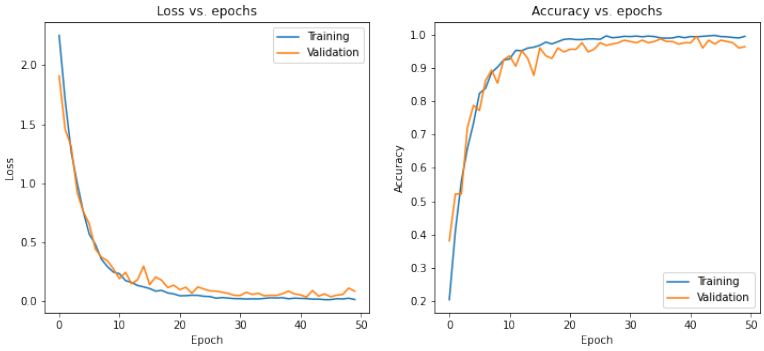

Рис. 4: Графики функции потерь и точности в течение обучения на all_male_speakers

train_data	test_speaker	cnn_loss	mlp_loss	cnn_accuracy	mlp_accuracy
speaker1	speaker1	0.106	0.109	0.982	0.976
speaker1	speaker2	14.215	7.658	0.159	0.174
speaker1	speaker3	6.224	4.788	0.455	0.386
speaker1	speaker4	9.352	8.324	0.235	0.227
speaker1	speaker5	2.475	3.45	0.661	0.612
speaker1	speaker6	15.957	9.493	0.152	0.182
all_male_speakers	speaker1	0.113	0.105	0.976	0.97
all_male_speakers	speaker2	0.184	0.452	0.955	0.864
all_male_speakers	speaker3	0.248	0.219	0.977	0.977
all_male_speakers	speaker4	0.289	0.357	0.939	0.917
all_male_speakers	speaker5	0.112	0.186	0.958	0.964
all_male_speakers	speaker6	5.086	9.082	0.394	0.109
all_speakers	speaker1	0.09	0.072	0.976	0.976
all_speakers	speaker2	0.14	0.505	0.947	0.864
all_speakers	speaker3	0.265	0.218	0.955	0.977
all_speakers	speaker4	0.364	0.308	0.917	0.909
all_speakers	speaker5	0.044	0.035	0.994	0.988
all_speakers	speaker6	1.344	2.591	0.794	0.648

Таблица 1: Результаты вычислений

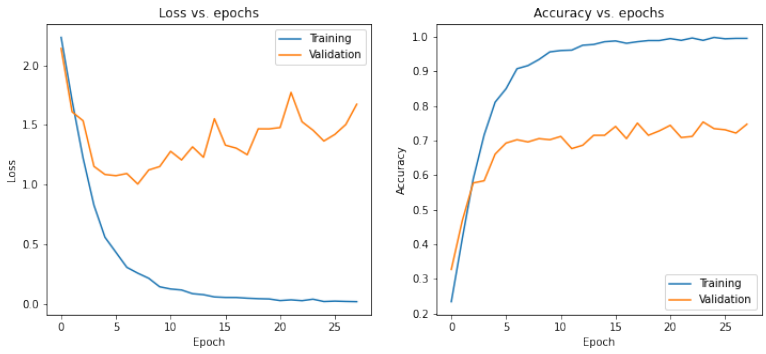

Рис. 5: Графики функции потерь и точности в течение обучения на all_speakers

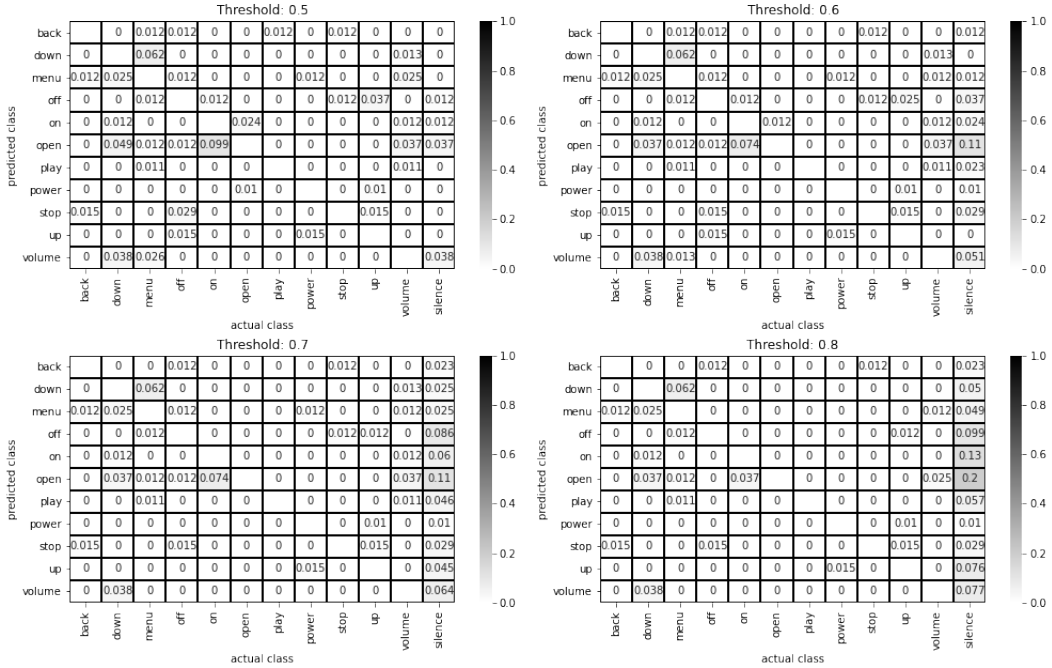

Рис. 6: Матрицы ошибок для случая обучения на all_speakers

Видно, что при обучении только на одном дикторе, распознавание на всех остальных работает плохо.

При обучении только на мужских голосах, распознавание на женском работает лучше, чем при обучении на одном, но все-равно очень плохо.

При обучении на всех голосах, распознавание на каждом голосе дает приемлемую точность. Однако стоит отметить, что если большая часть голосов в тренировочной части - мужские, то на женском голосе распознавание будет работать хуже, чем на мужских.

Выводы

Заключение

В данной работе:

- Проведена предобработка звуковых дорожек, содержащих команды в wav файлах
- Разработан алгоритм распознавания речевых команд
- Реализован алгоритм распознавания речевых команд
- Проведены вычислительные эксперименты, в результате которых показана работоспособность и эффективность работы алгоритма распознавания речевых команд.

Список использованных источников

[1] Davis K. N., Biddulph R., Balashek S. Automatic recognition of spoken digits // The Journal of the Acoustical Society of America, 1952. Vol. 24, No 6. P. 637-642

[2] Newell A. Harpy, production systems and human cognition // Research Showcase @ Carnegie Mellon University, 1978

[3] Plomp R., Pols L. C. W., van der Geer J.P. Dimensional analysis of vowel spectra // The Journal of the Acoustical Society of America, 1967, Vol. 41, P. 707-712

[4] Bogert B. P., Healy M. J. R., Tukey J. W. The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking // Proceedings of the Symposium on Time Series Analysis, 1963, Ch. 15, P. 209-243

[5] Davis S., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, Vol. ASSP-28, No. 4, P. 357-366

[6] Rabiner L.R., Sambur, M.R. An Algorithm for Determining the Endpoints of Isolated Utterances // The Bell System Technical Journal, 1975, Vol. 54, No. 2, P. 297-315

[7] Аксёнов О.Д. Метод мел-частотных кепстральных коэффициентов в задаче распознавания речи // 55-я юбилейная научная конференция аспирантов, магистрантов и студентов БГУИР, 2019, С. 45-46

[8] Rosenblatt, Frank. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

[9] LeCun Y., Bengio Y.. Convolutional networks for images, speech, and time-series // The Handbook of Brain Theory and Neural Networks, 1995, MIT

Приложение

Ссылка на репозиторий с программой веб-сервисом для записи датасета, состоящего из звуковых файлов: `https://gitlab.com/polotent/commandrecorder`

Ссылка на репозиторий с программой предобработки данных, обучением и тестированием нейронной сети: `https://gitlab.com/polotent/boxy`