УДК 004.52

Использование машинного обучения для распознавания речи

Цыбульский Анатолий Сергеевич
Студент. Российский экономический университет имени Г. В. Плеханова (Москва, Стремянный пер., 36)

Для решения повседневных бытовых и рабочих задач не всегда удобно пользоваться привычным средством ввода информации – клавиатурой. Перспективным для развития направлением ввода информации является управление голосом. Тем не менее, задачу распознавания человеческой речи нельзя считать полностью решенной – существующие системы распознавания активно развиваются, но имеют ряд недостатков, которые ведут к низкой точности распознавания человеческой речи - она уникальна и имеет непостоянную скорость и амплитуду.
В статье рассматривается возможность применения технологий машинного обучения для процесса распознавания человеческой речи. В статье подробно описаны все этапы процесса распознавания речи, включая этап подготовки звукового сигнала и финальный этап анализа результатов работы разработанного на основе рекуррентной нейронной сети программного приложения распознавания речи. Искусственные нейронные сети зарекомендовали себя в решении задач распознавания различных входных сигналов ввиду их способности к обобщению данных путем глубокого обучения, в особенности рекуррентные нейронные сети, которые в рамках каждой следующей итерации учитывают результаты предыдущих
Данная работа определяет потенциально перспективные области использования систем распознавания речи.
Ключевые слова: Распознавание речи, искусственные нейронные сети, Speech API

The use of machine learning for speech recognition

Tsybulskiy Anatoly Sergeevich
Student. Plekhanov Russian University of Economics (Moscow, Stremyaniy Street, 36)

To solve every day routine and work tasks it is not always convenient to use the usual mean of entering information - the keyboard. A promising direction for the development of information input is a voice control. Nevertheless, the task of human speech recognizing cannot be considered as completely solved one - recognition systems are actively developing nowadays, but still have a number of shortcomings that lead to low recognition accuracy of human speech - it is because any voice is unique and has a variable speed and amplitude.
Artificial neural networks have proven themselves in solving problems of recognizing different input signals because of their ability to generalize data through deep learning and training. The article deals with the possibility of applying machine learning technologies for human speech recognition.
The article contains description of developed speech recognition software complex that uses a recurrent artificial neural network. Potentially perspective areas for using speech recognition systems are presented.
Keywords: Human speech recognition, artificial neural networks, Speech API

Для решения повседневных бытовых и рабочих задач не всегда удобно пользоваться привычным средством ввода информации – клавиатурой. Перспективным для развития направлением ввода информации является управление голосом. Такой способ не требует наличия интерфейсов ввода, с которыми пришлось бы взаимодействовать вручную. Однако, голосовое управление больше всего зависит от технологий распознавания человеческой речи – диктованный текст или голосовые команды должны восприниматься автоматизированными системами корректно и полностью. От этого зависит достижение системами ожидаемого от них результата.

Преобразовывать примитивный речевой сигнал в цифровой вид, а затем распознавать его в текстовые данные научились еще в 1952 году [1]. Спустя полвека существует большое количество автоматизированных систем распознавания человеческой речи, приспособленных для решения самых разнообразных задач, среди которых обычно выделяют системы диктовки и системы приема команд. Тем не менее, задачу распознавания человеческой речи нельзя считать полностью решенной – существующие системы распознавания активно развиваются, но имеют ряд недостатков. Среди самых крупных недостатков выделяют, во-первых, неточность распознавания вовне тщательно контролируемой среды – в обычных помещения и оживленных улицах, а во-вторых при подачи вводной речевой информации разными людьми. В первом случае речевой сигнал может смешиваться с акустическим шумом окружающей среды, что сильно усложнит его анализ. Во втором случае, принимая во внимание амплитудно-частотные признаки речевого сигнала - одна и та же фраза, но произнесённая разными людьми, представляет собой два непохожих друг на друга сигнала [3].

Без решения этих недостатков невозможно представить полноценного применения голосового управления или диктовки в работе, быту. Это создает потребность в улучшении сервисов распознавания речи - необходимо сделать такие сервисы точными и быстрыми, способными выполнять задачу распознавания речи так же хорошо, как с этим справляется взрослый человек.

Таким образом, задача распознавания человеческой речи все еще остается актуальной проблемой. В данной статье будет рассмотрена возможность применения технологий машинного обучения для процесса распознавания человеческой речи и приведены потенциально перспективные области использования систем распознавания речи.

По определению, машинное обучение – это процесс, в ходе которого программа обучается корректному выполнению действий для достижения результата, а затем самостоятельно решает некоторые задачи, опираясь на свое обучение. Под обучением программы в контексте данной статьи понимается обработка большого числа примеров записей человеческого голоса - речей и разговоров. Программа в процессе обучения должна выявить закономерности в примерах и далее использовать их для распознавания характеристик новых для нее данных, лишь частично похожих на примеры из ее обучения. Машинное обучение реализуется набором из разнообразных алгоритмов, но одним из самых перспективных направлений в решении задачи распознавания человеческой речи является применение искусственных нейронных сетей. Искусственные нейронные сети зарекомендовали себя в решении задач распознавания различных входных сигналов ввиду их способности к обобщению данных путем глубокого обучения. Обучаясь, программа должна анализировать, что объединяет, а что противопоставляет между собой примеры, полученные в качестве входных данных.

Каждый раз, когда программа дает на выходе ответ на предложенный ей пример, обратно она получает вердикт: корректен ответ или нет. Для этого заранее перед обучением подготавливаются обучающая выборка – это примеры уже с заранее известными правильными и неправильными вариантами ответа. Применяя такой подход к обучению распознаванию речи – потребуются записи человеческой речи и текстовое сопровождение звукового ряда. Если программа в своем вердикте ошиблась, ее алгоритм по команде изменяет набор параметров и закономерностей, который программа использует в процессе распознавания речи.

В течение всего процесса обучения закономерности меняются столько раз, сколько необходимо, до тех пор, пока программа не будет справляться с примерами с определенным ожидаемым процентом ошибки. В конце обучения полученный набор параметров и закономерностей может использоваться для распознавания новых и неизвестных программе примеров речи. Целый процесс распознавания речи можно разбить на следующие этапы:

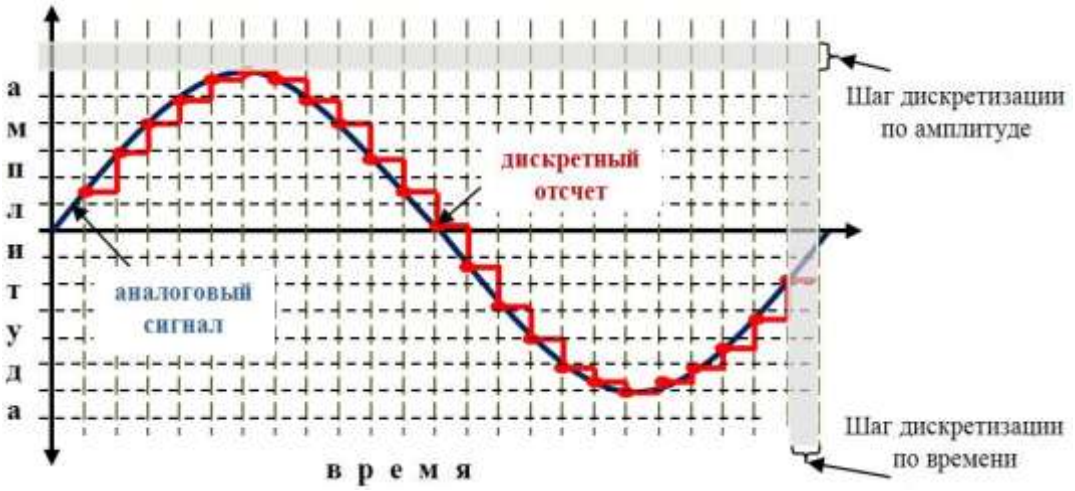
1. Передача звуков в распознающее программное приложение;
2. Обработка полученного звукового сигнала, оценка качества и определение помех;
3. Выделение участков сигнала, содержащих речь;
4. Распознавание речи в выделенных участках, в данном случае с использованием моделей машинного обучения.

Очевидно, что передавать речь в программное приложение распознавания необходимо в виде массива чисел. Этот процесс осложняется тем, что в отличие от входных сигналов другого формата (изображения, текст), звук в физическом мире представляет собой волны. Более того, человеческая речь имеет непостоянную скорость и амплитуду. Одна и та же фраза может быть произнесена быстро или медленно, что на выходе дает различные по размеру звуковые файлы, но содержащие одинаковую смысловую нагрузку.

Звуковые волны одномерны и на конечных отрезках непрерывны; это означает, что в каждый момент времени у них есть одно значение и оно зависит от амплитуды волны. Для преобразования звуковой волны в числовой формат, необходимо использовать значения амплитуды волны в равностоящих точках – звуковой сигнал необходимо *дискретизировать* [3].

Значения амплитуды волны записываются десятки тысяч раз в секунду. Согласно теореме Котельникова, для воссоздания исходной звуковой волны, почти идеально приближенной к исходной, достаточно использовать частоту дискретизации, вдвое превышающую самую высокую частоту записываемого звука. Эта доказанная теорема дает ответ на вопрос о необходимом и достаточном значении частоты дискретизации, и формально звучит как:

любую функцию, состоящую из частот от 0 до f_1, можно непрерывно передавать с любой точностью при помощи чисел, следующих друг за другом через $\frac{1}{(2*f_1)}$ секунд. Для качественного покрытия диапазона частот, которые человек способен различать на слух (20 Гц — 20 КГц) принято использовать частоту 44 100 Гц, что делает интервалы амплитуды равными 1/44100 секунды.



Прежде чем приступать к использованию моделей машинного обучения, необходимо оценить качество полученного, уже цифрового сигнала, затем из временной функции сигнала преобразовать в функцию частоты, тем самым облегчить работу алгоритмам распознавания. Цифровой сигнал можно сгруппировать на короткие фрагменты, длинной по несколько десятков миллисекунд. Далее, используя математическую операцию «преобразование Фурье», фрагменты разложим на элементарные гармонические составные части, начиная от самых нижних частот, и после этого можно просуммировать коэффициенты мощности частоты, чтобы получилась целостная частотная картина звуковой волны. В таких подготовленных данных нейронной сети будет проще находить шаблоны и закономерности, чем в сырых записях человеческой речи.

Теперь, когда речь представлена в цифровом формате, можно обучать глубокую искусственную нейронную сеть. Искусственная нейронная сеть – это совокупность соединенных и взаимосвязанных между собой искусственных нейронов, обрабатывающих входные значения и генерирующих выходной сигнал. Входные значения искусственные нейроны суммируют и применяют к ним функцию активации. Активность одного искусственного нейрона можно описать математической формулой $Y_j = F(\sum W_{ij} X_i)$, где Y_j – выходной сигнал j-нейрона, F – функция активации, W_{ij} – вес связи между нейронами i и j, X_i – выходной сигнал i-нейрона [4]. Топология самой нейронной сети определяется тем, как соединены в ней нейроны. По структуре связей выделяют сети прямого распространения и рекуррентные нейронные сети. В искусственных нейронных сетях прямого распространения

на каждый нейрон одного слоя подаются все выходные сигналы предыдущего слоя, что означает однонаправленную связь между слоями статической сети. В свою очередь в рекуррентных сетях присутствуют обратные связи, что делает выходной сигнал нейрона зависящим от его предыдущего состояния, а сама сеть является динамической.

Таким образом, искусственная нейронная сеть будет получать на вход фрагменты записей человеческой речи длиной в несколько миллисекунд; для каждого полученного фрагмента сеть должна определить, какой звук был произнесен. От сети на выходе ожидается набор вероятностей, отражающий уверенность сети в том, какой из известных ей звуков или фонем был произнесен. В ответ сеть получит вердикт, верно ли был распознан фрагмент. Программное приложение распознавания речи сравнивает входной аудио-фрагмент с репрезентативной акустической базой, которая, в идеале, должна содержать тысячи часов разнообразных надиктованных записей речей или телефонных разговоров. Таким образом реализуется принцип обучения сети «с учителем».

Непосредственно в рамках данной работы используется рекуррентная нейронная сеть, которая в рамках каждой следующей итерации учитывает результаты предыдущих. В качестве функции активации использовалась сигмоида, от характера плавности функции которой зависит степень нелинейности процесса обработки данных искусственным нейроном. Выбор такой топологии сети обосновывается тем, что каждая буква, входящая в слово, в реальном мире должна влиять на вероятную следующую букву в слове – такова особенность всех человеческих языков [2]. К примеру, если были произнесены буквы «П», «Р», «И», «В», то вероятнее всего, следующими произнесенными буквами будут «Е» и «Т», вместе образующие слово «Привет». И наоборот, очень маловероятно, что следующими произнесенными буквами могли бы быть «Ж» или «Щ».

Таким образом, запоминая предыдущие результаты, рекуррентная нейронная сеть сможет давать некоторые прогнозы о будущих выходных сигналах. Обрабатывая полученный сигнал фразы с самого начала, сеть не только распознает каждый звук или фонему, но и может с определенной точностью предсказывать конец фразы.

В свою очередь, для задач, где входящие сигналы между собой не связаны (например, диктовка цифр) – лучше использовать сеть прямого распространения сигнала. Результатом обучения сети является обученная модель, которая готова к распознаванию речи. А результатом работы сети над аудио фрагментом является разложение звукового ряда на наиболее вероятно произнесенные звуки.

На рисунке ниже представлена модель-схема алгоритма работы программного приложения с поясняющими комментариями:

Рисунок 1 - Схема работы программного приложения

Учитывая, что буквы определяются сетью по одной, нейронная сеть может предлагать на выходе свою уверенность в распознанных, но непроизносимых сочетаниях фонем. Поэтому, чтобы «очеловечить» результаты работы комплекса распознавания и отбросить нереалистичные сочетания фонем, необходимо уточнять результаты, сравнивая их с акустической базой письменного текста. Под такой базой может использоваться платформа компании Google - Cloud Speech API. Платформа имеет открытый для независимых разработчиков программный интерфейс облачного приложения и содержит в своей базе коллекции для 110 человеческих языков.

В рамках данной работы было подготовлено программное приложение, в состав которого входят:

- устройство приема голосовой информации на аппаратном уровне, передающее сигнал в программный модуль распознавания речи;
- программный модуль распознавания речи, использующий рекуррентную нейронную сеть для распознавания произнесенных фонем. Модуль написан с помощью высокоуровневого языка программирования Python 3.6;
- программный модуль обработки, взаимодействующий с интерфейсом приложения облачного сервиса Google Speech API, необходимый для обработки и уточнения результатов работы модуля распознавания речи.

Программное приложение реализовано в виде консольного приложения и позволяет начинать новую сессию распознавания по команде, а также записывать результат работы комплекса в текстовый файл.

```
def sigmoid(x):
    return 1 / (np.exp(-x) + 1)
def sigmoid_dgrad(x):
    return sigmoid(x) * (1 - sigmoid(x))
def neural_network(W, x):
    fordot_activation = sigmoid(np.dot(x, W[0]))
    return np.dot(fordot_activation, W[1])
def do_neural_network_reccur(W, x, k=1):
    return np.dot(np.dot(W[1].T, W[0].T**k), sigmoid_dgrad(x))
def reccur_loss_funct(W, x):
    loss_sum = 0.
    for x_i in x:
        out_net = neural_network(W, xi)[0][0]
        outplusstep = 1. + x_i * net_out
        d_net_out = do_neural_network_reccur(W, xi)[0][0]
        d_outplusstep = net_out + x_i * d_net_out
        func = f(x_i, outplusstep)
        errdouble = (d_outplusstep - func)**2
        summtoloss += errdouble
    return summtoloss
weight = [npr.randn(1, 10), npr.randn(10, 1)]
dltlm = 0.0001
for i in range(4000):
    totaloss =  grad(reccur_loss_funct)(W, x)
    weight[0] = weight[0] - dltlm * totaloss[0]
    weight[1] = weight[1] - dltlm * totaloss[1]
```

Рисунок 2 - фрагмент программного кода модуля нейронной сети

```
r = sr.Recognizer()
with sr.Microphone() as source:
    print("Начало записи")
    audio = r.listen(source)
```

```
Python 3.6.3 Shell
File  Edit  Shell  Debug  Options  Window  Help
= RESTART: C:/Users
Начало записи
было произнесено:
тест распознавания голоса 123
конец записи
= RESTART: C:/Users
Начало записи
было произнесено:
Съешь ещё этих мягких французских булок да выпей же чаю
конец записи
= RESTART: C:/Users
Начало записи
было произнесено:
123456789 10
конец записи
```

Рисунок 3 - Результат работы программного приложения

Основанные на проприетарном программном обеспечении технологии распознавания человеческой речи уже применяются в большинстве сервисов ведущих мировых IT-компаний: Google, Apple, Яндекс, Microsoft, Abbyy и так далее. Среди самых популярных вариантов использования технологий распознавания речи стоит выделить сайты онлайн-перевода, поисковые и навигационные системы, встроенные службы распознавания голосовых команд в операционных системах Windows, MAC OS/iOS, Android. Сферами, где применение распознавания человеческой речи будет очень перспективным и будет нести заметную пользу, можно назвать – муниципальные жилищно-коммунальные хозяйства и концепцию «умного дома», где распознавание голосовых команд позволит усовершенствовать функциональные возможности отечественных информационных систем в сфере ЖКХ [5] в условиях возможной «цифровизации» ЖКХ в соответствии с принятой в июле 2017 года программой «Цифровая экономика Российской Федерации» [6].

Заключение

В данной работе были проанализированы недостатки существующих систем распознавания речи, изучена возможность применения технологий машинного обучения для улучшения этого процесса, включая этапы предварительного преобразования звукового сигнала в цифровой формат. В работе представлен способ решения задачи распознавания человеческой речи с использованием технологии машинного обучения на основе рекуррентной нейронной сети, с использованием внешней фонетико-акустической базы и применяющий предварительную обработку входного звукового сигнала. Дано описание топологии и характеристик используемой нейронной сети. По итогам приведены потенциально перспективные области использования систем распознавания речи, использующих технологии машинного обучения.

Описанная модель решения задачи применяется для отдельных слов русского языка. Точность распознавания в контролируемых условиях (без посторонних шумов, с четким произношением слов) достигает 90%. Погрешность в результатах распознавания текста в условиях естественного шумового фона и произвольных помех составляет до 20%, что можно считать приемлемым в рамках конечного времени работы программы на рабочей станции.

В течение оценки работы решения, выявлена зависимость качества результатов в самой большей степени от топологии сети и количества искусственных нейронов. Как следствие, подготовленное решение может быть модернизировано для повышения точности распознавания путем усовершенствования топологии нейронной сети и проведения экспериментов с количеством искусственных нейронов.

Список литературы

1. Брагина Е. К., Соколов С. С. /Современные методы биометрической аутентификации: обзор, анализ и определение перспектив развития // Вестник АГТУ. 2016. №61.
2. Комарцова Л. Г., Максимов / А. В. Нейрокомпьютеры. 1 изд. // М.: Изд-во МГТУ им. Н.Э. Баумана, 2002.
3. Попов А.А., Телемтаев М.М. Совершенствование отечественных информационных систем управления недвижимостью на основе зарубежного опыта // Прикладная информатика. 2012, № 2, 18-25.
4. Тампель И.Б., Карпов А.А. / Автоматическое распознавание речи. Учебное пособие// Санкт-Петербург: СПб: Университет ИТМО, 2016.
5. Цифровая экономика // Минкомсвязь России URL: http://minsvyaz.ru/ru/activity/directions/779/ (дата обращения: 20.12.2017).
6. Biddulph, Davies / Davies Automatic Speech Recognition of Spoken Digits // The Journal of the Acoustical Society of America. – 1952. – 6. – С. 637-642.