электронный журнал
МОЛОДЕЖНЫЙ
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО «Московский государственный технический университет им. Н.Э. Баумана»

Распознавание голосовых команд с помощью самоорганизующейся нейронной сети Кохонена

Кладов Станислав Александрович,
ИУ7-83. МГТУ им. Н.Э. Баумана
stas@kladov.ru

Широкое развитие роботов-манипуляторов и нейронных сетей за последнее время диктует необходимость создания методов и алгоритмов поддержки принятия решений, в том числе основанные на использовании нейронных сетей, гибридных моделей и математического аппарата нечеткой логики. В этой связи важным направлением становится распознавание голосовых команд.

Одними из известнейших поставщиков ПО в этой области являются Google (Google Voice Search) и Apple (Siri). На отечественном рынке речевыми технологиями в основном занимается Центр Речевых Технологий.

В данной статье рассмотрен способ реализации простейшей системы распознавания голосовых команд с помощью нейронной сети Кохонена.

Характеристики систем распознавания речи

В настоящий момент системы распознавания речи характеризуются следующими признаками:

- дикторозависимость;
- раздельность речи;
- назначение.

Дикторозависимая система предназначена для использования одним диктором, в то время как дикторонезависимая система предназначена для работы с любым диктором. Дикторонезависимость – очень ценное качество системы, но в то же время очень труднодостижимое, так как при обучении системы она настраивается на параметры того диктора, на примере которого обучается. Таким образом, в процессе создания дикторонезависимой системы применяются гораздо более сложные алгоритмы обучения. В таких системах частота ошибок распознавания обычно в 3-5 раз больше, чем в дикторозависимых.

Если в речи слова разделяются интервалами тишины, то говорят, что эта речь – раздельная. Естественная речь, как правило, слитная. Распознавание слитной речи намного труднее в связи с тем, что границы отдельных слов не четко определены и их произношение сильно искажено смазыванием произносимых звуков.[4]

Назначение системы определяет требуемый уровень абстракции, с которым будет происходить распознавание произнесенной речи. Можно выделить 2 типа систем распознавания речи:

- командные системы;
- системы диктовки.

В командных системах, в общем случае, распознавание слова или фразы происходит как распознавание единого речевого элемента. То есть, при распознавании учитываются исключительно физические характеристики сигнала, а не смысловая нагрузка произнесенной речи.

Системы диктовки также анализируют контекст речевого элемента и поэтому требуют большей точности распознавания. Алгоритмы, задействованные в таких системах, например, скрытые сети Маркова, анализируют не только уникальные параметры самого речевого сигнала, но и контекст каждого произнесенного речевого элемента. Также могут применяться алгоритмы динамического программирования. Для анализа контекста в системе необходимо предусмотреть набор грамматических правил, которым должен удовлетворять произносимый и распознаваемый текст. Чем строже эти правила, тем проще реализовать систему распознавания, и тем ограниченней будет набор предложений, которые она сможет распознать [4].

Этапы создания системы

В качестве примера рассмотрим алгоритм создания командной дикторозависимой системы, который можно разделить на несколько шагов:

- получение сигнала;
- обработка сигнала;
- выделение уникальных параметров сигнала;
- обработка параметров нейросетью.

Для создания нейронной сети Кохонена следует выделить следующие шаги:

- проектирование топологии;
- обучение.

Проектирование нейронной сети Кохонена

Топология сети

В качестве классификатора в проектируемой системе выбрана самоорганизующаяся сеть Кохонена ввиду простоты ее топологии и алгоритма обучения.

Сеть Кохонена характеризуется тем, что состоит из слоя входов и связанного с ним единственного слоя нейронов, отвечающих каждый за свой класс команд. На слой входов подается вектор параметров, которые затем анализируются каждым нейроном на принадлежность к соответствующему классу.

Предположим, что проектируемая система должна распознавать 5 команд:

«Старт», «Стоп», «Красный», «Синий», «Зеленый».

Топология рассматриваемой сети Кохонена будет выглядеть следующим образом (см. рисунок 1):

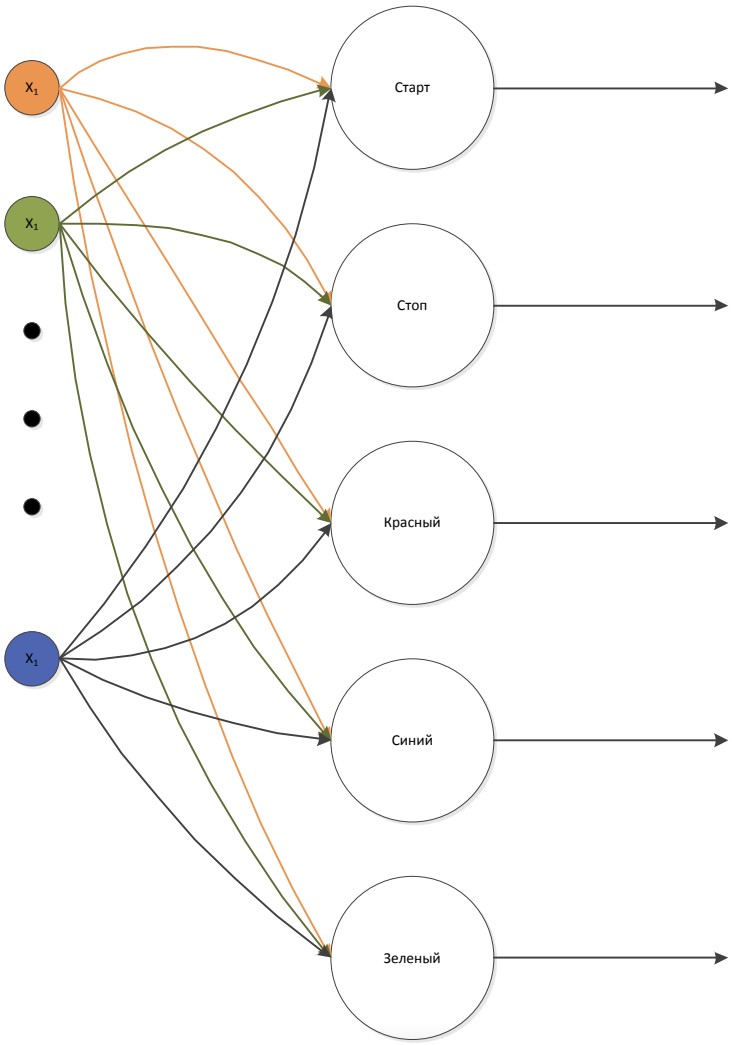

Рисунок 1. Топология нейросети.

Обучение нейросети

Слой Кохонена классифицирует входные векторы в группы схожих. Это достигается с помощью такой подстройки весов слоя Кохонена, что близкие входные векторы активируют один и тот же нейрон данного слоя.

Обучение Кохонена является самообучением, протекающим без учителя. Поэтому трудно (и не нужно) предсказывать, какой именно нейрон Кохонена будет активироваться для заданного входного вектора. Необходимо лишь гарантировать, чтобы в результате обучения разделялись несхожие входные векторы.[2]

Уравнение, описывающее процесс обучения, имеет следующий вид:

$$w_н = w_с + \alpha(x - w_с),$$

где $w_н$ – новое значение веса, соединяющего входную компоненту x с выигравшим нейроном; $w_с$ – предыдущее значение этого веса; α – коэффициент скорости обучения, который может варьироваться в процессе обучения.[3]

Таким образом, для обучения рассматриваемой системы необходимо подготовить обучающую выборку, которая содержит различные варианты произношения всех необходимых команд.

Получение входного сигнала

Входной сигнал — поток звуковых данных, записанный с высокой дискретизацией (20 КГц при записи с микрофона либо 8 КГц при записи с телефонной линии).[5]

Обработка полученного сигнала

Для повышения точности определения команды необходимо в поступающем сигнале отделить активную речь от фонового шума. Одним из наиболее известных алгоритмов отделения шума является VAD (Voice Active Detection), который основывается на кодировании пауз.

Анализ сигнала

Целью этого шага является получение параметров обработанного сигнала. Рассмотрим несколько основных алгоритмов.

Спектральный анализ (Анализ Фурье)

Спектр звукового сигнала является одним из важнейших инструментов анализа и обработки звука. Цель спектрального анализа — разложить ряд на функции синусов и косинусов различных частот, для определения тех, появление которых особенно существенно и значимо.

Ряд Фурье — представление произвольной функции f с периодом τ в виде ряда.

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{+\infty} A_k \cos\left(2\pi \frac{k}{\tau} x + \theta_k\right)$$

A_k — амплитуда k-го гармонического колебания,

$2\pi \frac{k}{\tau}$ — круговая частота гармонического колебания,

θ_k — начальная фаза k-го колебания.

a_0 – коэффициент Фурье функции f.

На рисунке 2 приведен пример разложения сигнала на гармоники.

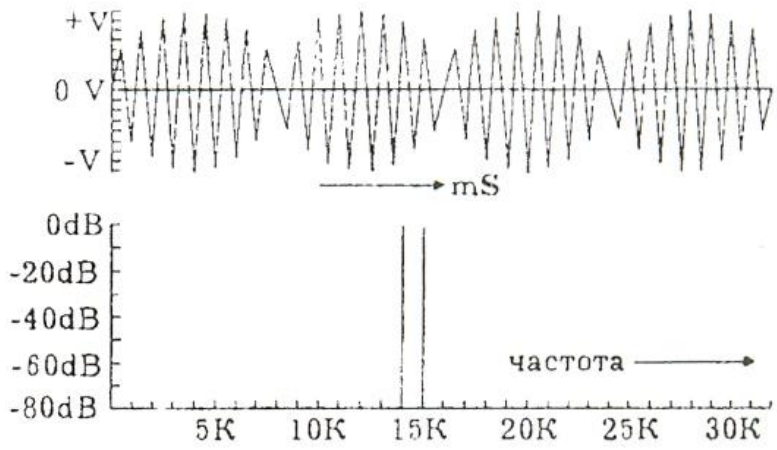

Рисунок 2. Пример спектра сигнала.

Поскольку количество входов в спроектированную сеть ограничено, то фиксирование количества гармоник может привести к потере точности распознавания длинных команд.

Кепстральный анализ

Этот метод оперирует понятием «кепстр». Кепстр — это спектр логарифма спектра исходного сигнала, т.е. первоначальный спектр представлен в логарифмическом масштабе. Таким образом, возможно представить исходную спектральную информацию еще более компактно, когда каждый гармонический ряд исходного спектра будет представлен всего одной составляющей в кепстре.[1]

Растрирование спектра

Другой метод анализа сигнала и выделения параметров – растрирование. Спектр — это наглядное, компактное, численное представление периодичностей, присутствующих в сигнале во временной области. Сигнал в частотной области представляется одной точкой, координаты которой по осям "X" и "Y" содержат информацию о частоте (или периоде сигнала) и амплитуде. При растрировании пиксели, формирующие изображение спектра сигнала, подаются на вход нейронной сети.

Спроектированную систему распознавания речи можно использовать для управления агентом (рациональным, интеллектуальным и т.д.), например, с помощью нечеткого блока принятия решений (см. рисунок 3).

Рисунок 3. Использование системы распознавания речи для управления агентом.

Система распознавания голосовых команд, реализованная с учетом приведенных выше методов, будет дикторозависимой. Однако, представляется возможным её усовершенствовать до дикторонезависимой, дополнив обучающую выборку вариантами произношения допустимых команд различными дикторами.

Список литературы

1. В.Н. Суворов «О кепстральном анализе в популярной форме», 2006 URL: http://robotics.bstu.by/mwiki/images/c/ca/PopularCepstral.pdf (дата обращения 29.02.12).
2. Л.Г. Комарцова, А.В. Максимов, Нейрокомпьютеры, М., Издательство МГТУ им. Н.Э. Баумана, 2004, 399 с.
3. Ф. Уоссермен, Нейрокомпьютерная техника, 1992, М., Мир, 1992, 184 с.
4. Распознавание речи. Часть 1. Классификация систем распознавания речи

URL: http://habrahabr.ru/blogs/artificial_intelligence/64572/ (дата обращения 29.02.12).
5. Распознавание речи. Часть 2. Типичная структура системы распознавания речи
URL: http://habrahabr.ru/blogs/artificial_intelligence/64594/ (дата обращения 29.02.12).