ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ В ОБЛАСТИ ИДЕНТИФИКАЦИИ ГОЛОСА

Н.И. Данков, *магистрант МТУСИ, 111024, г. Москва, ул. Авиамоторная, 8А, ndankov@list.ru*

УДК 004.896

Аннотация. В данной работе была исследована возможность применения нейронных сетей для идентификации речи. В частности, были рассмотрены стандартные подходы к распознаванию речи, определено понятие искусственного нейрона, как объекта, используемого в идентификации речи. Был исследован вариант распознавания речи с помощью нейронной сети, и представлены шаги для выполнения этой задачи.

Ключевые слова: нейронная сеть; распознавание речи; искусственный интеллект; машинное обучение.

RESEARCH OF THE POSSIBILITIES OF NEURAL NETWORK TECHNOLOGIES IN THE FIELD OF VOICE IDENTIFICATION

Nikita Dankov, *master's student MTUCI, 111024, Moscow, ul. Aviamotornaya, 8A*

Annotation. In this article the possibility of using neural networks for speech identification was investigated. In particular, standard approaches to speech recognition were considered, the concept of an artificial neuron as an object used in speech identification was defined. A speech recognition option using a neural network was investigated, and steps were presented to accomplish this task.

Keywords: neural network; speech recognition; artificial intelligence; machine learning.

Распознавание речи является альтернативой традиционным методам взаимодействия с компьютером, например, текстовому вводу через клавиатуру. Эффективная система распознавания может заменить или снизить вероятность использования стандартной клавиатуры и мыши. Система распознавания речи состоит, обычно, из четырех составляющих:

1. Микрофон, в который человек должен говорить.
2. Программное обеспечение распознавания речи.
3. Компьютер, на котором можно обрабатывать и интерпретировать речь.
4. Звуковая карта.

Основные определения, описывающие параметры речи человека, а также связанные с формой, размерами, динамикой трансформации образующего ее тембра и эмоциями (эмоциональным состоянием) человека, делятся на четыре группы признаков, позволяющих объективно различать речевые сигналы: спектрально-временные, амплитудно-частотные, кепстральные и последние – нелинейной динамики.

По типу речи различаются системы распознавания слитной речи и речевых сигналов. В последнем случае требуется дискретное (специальное) произнесение речевых команд, где паузы между ними значительно больше внутрисловных пауз. Обычно длительность таких разделительных пауз составляет половину секунды.

В ходе распознавания слитной речи слова фраз произносятся естественно, без вставки между словами каких-либо специальных пауз. Существует также и третий вариант работы систем распознавания, где они должны обнаруживать произношение в звуковом потоке заданных слов, независимо от «зашумления» другими словами или выделения паузами. Данный режим называется поиском ключевых слов [1-3].

В основном используют два варианта в распознавании речи:

1. Мера близости параметров.
2. Распознавание с помощью нейронных сетей.

Нейронные сети не делают предположений о статистических свойствах объектов и имеют несколько качеств, что делает их привлекательными моделями для распознавания речи. При использовании для оценки вероятности сегмента речи нейронные сети позволяют проводить дискриминационную тренировку естественным и эффективным образом. Мало предположений о статистике входных функций сделаны с нейронными сетями. Однако, несмотря на их эффективность в классификации краткосрочных временных единиц, таких как отдельные фонемы и отдельные слова, нейронные сети редко бывают удачными для непрерывных задач распознавания, в основном из-за отсутствия способности моделировать временные зависимости. Вариант глубокого обучения нейронных сетей был использован в экспериментах для решения этой проблемы.

Из-за неспособности исходных нейронных сетей к моделированию временных зависимостей альтернативный подход заключается в использовании нейронных сетей в качестве предварительной обработки, например, для преобразования признаков, уменьшения размерности [2, 5].

Другой подход – нейронная сеть с временной задержкой. Он использовал модифицированный вариант обучения для захвата пространственных отклонений и временных деформаций в последовательности функции. Один слой ввода, два скрытых слоя и один выходной слой были использованы для классификации различных фонем, созданных носителями языка. Весовые коэффициенты были определены таким образом, что система была несколько инвариантна к временным искажениям в речевом сигнале. Он признавал только речь на фонеме и не использовался для решений в более длительные промежутки времени, т. е. он не использовался непосредственно для распознавания слов.

Слоги и слова по существу являются последовательными. Это означает, что обе методики очень сильны в другом контексте. Как и в нейронной сети, задача состоит в том, чтобы установить соответствующие веса соединения, задача марковской модели – найти соответствующие вероятности перехода и наблюдения. Во многих системах распознавания речи оба метода реализуются вместе и работают в симбиотических отношениях. Нейронные сети очень хорошо справляются с изучением вероятности фонемы из высокопараллельного аудиовхода, в то время как модели Маркова могут использовать вероятности наблюдения фонем, которые предоставляют нейронные сети для получения наиболее вероятной последовательности или слова фонемы. Это лежит в основе гибридного подхода к пониманию естественного языка [4].

Стандартная структура распознавание речи представлена на рис. 1.

Рисунок 1

Первый шаг, состоит из акустического окружения, в том числе из оборудования для принятия речи. Это окружение может оказать сильное влияние на сгенерированные речевые представления. Например, она может оказывать дополнительное влияние, возникающее в результате аддитивного шума или комнатная вибрации.

Второй шаг предназначен для решения этих проблем, а также получения акустических представлений, в которых необходимо разделять классы речевых звуков и эффективно подавлять нерелевантные источники вариации.

На третьем шаге должны быть выделены специфические особенности предварительно обработанного сигнала. Это можно сделать с использованием различных методов, таких как кепструм-анализ и спектрограмма.

Четвертый шаг классифицирует извлеченные функции и связывает входной звук с наилучшим подходящим звуком в известный «набор слов» и представляет это как результат.

Важно подавать нейронную сеть с нормализованным вводом. Записанные образцы никогда не воспроизводят идентичные формы сигнала; длина, амплитуда, фоновый шум могут различаться. Поэтому нам нужно выполнить предварительную обработку сигнала, чтобы извлекать только информацию, связанную с речью. Это означает, что использование правильных функций имеет решающее значение для успешной классификации. Хорошие функции упрощают дизайн классификатора, тогда как слабые функции (с небольшой степенью дискриминации) вряд ли могут быть скомпенсированы любым классификатором.

На первом этапе важно правильно отфильтровать шумы для того, чтобы получить правильный спектр. Входной сигнал можно очистить по специальному фильтру, заданному по формуле 1.

$$X_i = (X_i - 0{,}9 * X_{i-1})[0{,}54 - 0{,}46 * \cos((i-6) * \frac{2*\pi}{180})] \qquad (1)$$

где: $X_i -$ входные звуковые значения

На следующем шаге обычно необходимо получить правильную спектограмму, которую можно будет в дальнейшем использовать. Для этой задачи подойдет метод преобразования Фурье (формула 2).

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots . N-1 \qquad (2)$$

где:

N – длина кадра;

x_n – амплитуда сигнала;

X_k – комплексная амплитуда сигналов.

Можно выделить два основных подхода:

1. Подход с фиксированной точкой.

Этот подход переводит изменяющуюся во времени траекторию в точку в пространстве, уменьшая проблему классификации траекторий в одноточечную классификацию. Поскольку высказывания демонстрировали значительную вариативность, не только в функции, но и во времени, они были нормализованы так, что все временные деформации были отменены. Все траектории на тренировочных и испытательных множествах были нормированы на траектории 100 точек, так что евклидово расстояние между соседними точками в приведенном пространстве объектов всегда составляло одну сотую длины траектории, проецируемой в уменьшенное пространство.

2. Подход с разной траекторией.

Этот подход, вместо того чтобы сводить проблему к точечной классификации, имеет прямое отношение к нормализованным траекториям. Траектории непосредственно подаются в распознаватель, который в то же время непрерывно производит выход, который может быть использован для целей классификации.

Хотя ни один из подходов не был достаточно хорош для практических целей с нынешней степенью развития, они были достаточно хороши, чтобы доказать, что перевод речи в

траектории в пространстве объектов работает для целей признания. Человеческая речь – это динамический процесс, который можно правильно описать как траекторию в определенном пространстве признаков. Более того, схема уменьшения размерности доказала уменьшение размерности при сохранении некоторой исходной топологии траекторий, т.е. она сохранила достаточно информации, чтобы обеспечить хорошую точность распознавания. Интересно отметить, что, несмотря на то, что этот подход использовался в области распознавания речи более десятилетия, никто не использовал его для создания траекторий, а только для генерации последовательностей меток. Наконец, новый подход, разработанный для обучения архитектуре нейронной сети, оказался простым и очень эффективным. Это значительно сократило количество вычислений, необходимых для нахождения правильного набора параметров [2, 4].

При распознавании наиболее часто используется нейронная сеть — многослойный персептрон. Его общая структура состоит из нескольких слоев. Нейроны в его структуре часто функционируют по модели МакКаллока-Питса, соответствующей следующей функции, показанной в виде формулы 3 [9].

$$y_k(t) = f(u_k(t)) \qquad (3)$$

Наиболее известным алгоритмом обучения для многослойного персептрона является процедура, которая была описана Фрэнком Розенблаттом в 1959 г., и как ее модификация, предложенная Дэвидом Румельхартом как алгоритм обратного распространения ошибки (*back propagation error*), который позволяет осуществить управляемое обучение (обучение «с учителем»).

Основным моментом при обучении является коррекция весов, она производится в по следующей целевой функции, показанной в формуле 4.

$$w_{ij}(t+1) - \eta \frac{\delta E}{\delta w_{ij}(t)} x_i(t) \qquad (4)$$

где:

$\eta -$ коэффициент обучения.

Также часто используются в задачах распознавания голоса сверточные нейронные сети.

Сверточная нейросеть является особым видом нейросетей прямого распространения. Под прямым распространением понимается то, что переменные нейроны в этой сети разбиты на группы, называемые слоями.

Главный момент заключается в создании свертки, которую можно описать следующей формулой 5.

$$(f * g)[m,n] = \sum_{k,l} f[m-k, n-l] * g[k,l] \qquad (5)$$

где:

f – исходная матрица.

Подход с нейронными сетями подразумевает для начала получение самих данных о речи.

После первого шага, когда у нас получен некий набор данных, мы можем построить таблицу (табл.1), в которой каждому элементу соответствует набор чисел [1].

Таблица 1

Сегмент	1-ый результат	2-ой результат	….	i-результат
1-ый сегмент	x_{11}	x_{12}	….	x_{1i}
2-ой сегмент	x_{21}	x_{22}	…..	x_{2i}
…			….	
j-сегмент	x_{N1}	x_{N2}	…..	x_{Ni}

Получается, что к каждому нейрону определяется набор значений. Но на последнем слое у нас будет всего лишь один нейрон.

Это так называемая нейронная сеть с обратной связью.

Таким образом, весь алгоритм можно структурировать следующим образом:

1. В начале определить для каждого сегмента набор значений.
2. На втором шаге первоначальные значения подать на вход системе, при соответствии формуле 6.

$$y_j = f(\sum_{i=1}^{I} w_{ij}x_{ij} + \beta_j + w_j x_j) \quad (6)$$

где:
f*(x)*– это нелинейная функция
3. Далее вычислить последний слой (формула 7)

$$y_k = f(\sum_{j=1}^{J} w_{jk}y_i + \beta_k) \quad (7)$$

В целом, общий алгоритм будет выглядеть следующим образом: производится пошаговое конструирование матрицы с весами и постепенное уменьшение вероятности получить ошибку.

Сам процесс обучения происходит следующим образом: на вход нейронной сети подается выборка и подстраиваются необходимые весы. Процесс повторяется до тех пор, пока уровень ошибок не достигнет минимума.

Функция ошибки вычисляется формулой 8.

$$E = \frac{1}{2N}\sum_{i=1}^{N}(y_{ki} - d_i)^2 \quad (8)$$

где:
N – общее количество подданного на вход нейронной сети выборок;

Таким образом, вышеупомянутая модель может стать основой для распознавания речи на уровне нейронных сетей.

Литература

1. Ле Н.В., Панченко Д. Распознавание речи на основе искусственных нейронных сетей [Текст] // Технические науки в России и за рубежом: материалы Междунар. науч. конф. (г. Москва, май 2011 г.). – М.: Ваш полиграфический партнер, 2011. – С. 8-11. – URL https://moluch.ru/conf/tech/archive/3/712/ (дата обращения: 17.10.2018).
2. Николенко С, Кадурин А., Архангельская Е. В., Глубокое обучение. Погружение в мир нейронных сетей. – Питер, 2018. – 481 с.
3. Aurelien Geron Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. – O'Reilly Media, 2017. – 572 p.
4. Daniel Jurafsky, James H. Martin Speech and Language Processing. – Prentice Hall, 2008. – 1032 p.
5. Stuart Russell Artificial Intelligence: A Modern Approach. – PE; 3rd edition, 2015. – 1164 p.
6. Рязанов В.В. Модели, методы, алгоритмы и архитектуры систем распознавания речи. – М.: Вычислительный центр им. А.А. Дородницына РАН, 2013.
7. Садыхов Р. Х., Ракуш В. В. Модели гауссовых смесей для верификации диктора по произвольной речи // Доклады БГУИР, Минск. 2003. – № 4. – С. 95-103.
8. Тампель И.Б. Автоматическое распознавание речи – основные этапы за 50 лет // Научно-технический вестник информационных технологий, механики и оптики, 2015. – Т. 15. – № 6. – С. 957-968.
9. Сагациян М.В. Разработка и исследование коллективных нейросетевых алгоритмов дикторонезависимого распознавания речевых сигналов: Диссертация … кандидата технических наук. – Владимир, 2015.
10. Dong Yu, Li Deng. Automatic Speech Recognition. A Deep Learning Approach. London, Springer, 2015, 321 p.
11. Hinton G., Deng L., Yu D., Dahl G., Mohamed A.-R., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the

shared views of four research groups. IEEE Signal Processing Magazine, 2012, vol. 29, no. 6, pp. 82-97.
12. Deng L. Deep learning: from speech recognition to language and multimodal processing // APSIPA Transactions on Signal and Information Processing, 2016. vol 5. pp. 1-15.
13. Кипяткова И.С., Карпов А.А. Разновидности глубоких искусственных нейронных сетей для систем распознавания речи // Труды СПИИРАН, 2016. Вып. 6 (49). – С. 80-103.
14. Меденников И.П. Методы, алгоритмы и программные средства распознавания русской телефонной спонтанной речи. – М.: Проспект, 2015. – 456 с.
15. Morioka T., Iwata T., Hori T., Kobayashi T. Multiscale recurrent neural network based language model // Proceedings of INTERSPEECH-2015. 2015. pp. 2366-2370.