УДК 004.032.26:004.85

ИССЛЕДОВАНИЕ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ НА ОСНОВЕ АКУСТИЧЕСКОГО И ЯЗЫКОВОГО МОДЕЛИРОВАНИЯ

Хлопенкова А.Ю., Белов Ю.С.

Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», Калуга, e-mail: annakhl@yandex.ru

Данная статья посвящена системе автоматического распознавания речи на основе различных алгоритмов. Выделяются и описываются такие алгоритмы, как Dynamic Time Warping (DTW), Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Deep Neural Network (DNN), Likelihood Ascent Search (LAS). Анализируется работа каждого из них. Рассматривается процесс преобразования речевого сигнала в набор чисел посредством сэмплирования. Описывается построение скрытых марковских моделей для различных типов фонем. Объясняется трёхслойная архитектура алгоритма глубоких нейронных сетей. Выделяются недостатки алгоритма нейронных сетей. На основе анализа делаются выводы о достоинствах и недостатках автоматического распознавания в целом, а также демонстрируется область применения систем распознавания речи. В заключение прогнозируется дальнейшая область развития распознавания речи и требования, которые будут предъявляться к новым методам.

Ключевые слова: распознавание речи, алгоритм, сэмплирование, нейронные сети, DTW, HMM, ANN, DNN, LAS

INVESTIGATION OF ALGORITHMS OF AUTOMATIC SPEECH RECOGNITION BASED ON ACKUSTIC AND LANGUAGE SIMULATION

Khlopenkova A.Yu., Belov Yu.S.

Kaluga branch of the federal state budget education institution of higher education «Moscow State Technical University named after N.E. Bauman (National Research University)», Kaluga, e-mail: annakhl@yandex.ru

This article is devoted to the system of automatic speech recognition, based on various algorithms. Such algorithms as Dynamic Time Warping (DTW), Hidden Markov Models (HMM), Artificial Neural Networks(ANN), Deep Neural Network (DNN), Likelihood Ascent Search (LAS) are identified and described. The work of each of them is analyzed. Considered the process of converting a speech signal into a set of numbers by means of sampling. Described the construction of hidden Markov models for different types of phonemes. Explained the three-layer architecture of the algorithm of deep neural networks. Deficiencies of the neural network algorithm are highlighted. Based on the analysis, conclusions are drawn about the advantages and disadvantages of automatic recognition in general and demonstrated the field of application of speech recognition systems. In conclusion, the further development of speech recognition is predicted and the requirements, that will be presented to new methods.

Keywords: speech recognition, algorithm, sampling, phonemes, neural networks, DTW, HMM, ANN, DNN, LAS

Распознавание речи – это возможность машины или программы идентифицировать слова и фразы на устном языке и преобразовать их в машиночитаемый формат [1]. Речь представляет собой последовательность звуков. Звук, в свою очередь, представляет собой суперпозицию звуковых волн разных частот. Волна, как известно из физики, характеризуется двумя атрибутами – амплитудой и скоростью. Чтобы сохранить аудиосигнал на цифровом носителе, его необходимо разделить на несколько промежутков и принять определенное «усредненное» значение для каждого из них. Таким образом, механические колебания преобразуются в набор чисел, подходящих для обработки на современных компьютерах. Рудиментарное программное обеспечение для распознавания речи имеет ограниченный словарный запас слов и фраз, и поэтому оно может идентифицировать слова, только если произношение очень четкое. Более сложное программное обеспечение имеет возможность принимать естественную речь.

Распознавание речи работает на основе двух алгоритмов: акустического и языкового моделирования. Акустическое моделирование представляет собой взаимосвязь между лингвистическими единицами речи и аудиосигналов; языковое моделирование соответствует звукам с последовательностями слов, чтобы помочь различать слова, которые звучат одинаково. Процесс автоматического преобразования речи в текст может быть представлен в виде выражения

$$W = \arg\max P(A \mid W) * P(W),$$

где argmax – значение аргумента, при котором выражение достигает максимума, $P(A|W)$ – вероятность появления гипотезы по оценке акустической модели при условии появления гипотезы по оценке языко-

вой модели, P(W) – вероятность появления гипотезы W по оценке языковой модели [1–2].

Речь захватывается чувствительным к звуку элементом в микрофоне, который преобразует переменное звуковое давление в эквивалентные изменения электрического сигнала, то есть тока или напряжения. Затем этот аналоговый сигнал отбирается и квантуется в цифровой бит-поток (формат). Далее происходит сэмплирование – процесс получения значений аналогового сигнала в отдельные моменты времени T, где квантование достигается путем преобразования амплитуды в каждый момент выборки в дискретное двоичное число с заданной длиной бит. Этот двухступенчатый процесс иногда называют модуляцией импульсного кода – PCM (Pulse Code Modulation). Количество выборок в секунду (частота) f_s в Гц равно обратному периоду выборки, то есть $f_s = 1/T$. Теорема выборки утверждает, что частота дискретизации должна быть как минимум в два раза выше самой высокой частотной составляющей, присутствующей в сигнале. Если используется меньшее количество образцов, возникает явление, известное как сглаживание, когда при повторной конструкции может появляться сигнал с более низкой частотой. Частота дискретизации для типичной речи приблизительно равна 3,3 кГц. Уже при 6–20 кГц требуется фильтр предварительной выборки или сглаживания, чтобы удалить частотные компоненты выше частоты Найквиста.

Производительность систем распознавания речи обычно оценивается с точки зрения точности и скорости. Точность оценивается как количество ошибок в слове, тогда как скорость измеряется с коэффициентом реального времени. Другие меры точности включают единичную ошибку и коэффициент успеха команды.

В процессе развития системы распознавания речи постепенно появлялись новые алгоритмы работы, такие как динамическое временное деформирование, скрытые марковские модели, нейронные сети и распознавание речи end-to-end [1, 3–4].

Dynamic Time Warping

Одним из самых ранних алгоритмов является алгоритм распознавания речи на основе динамического временного деформирования (DTW – Dynamic Time Warping). В анализе временных рядов динамическое временное деформирование является одним из алгоритмов для измерения сходства между двумя временными последовательностями. DTW применяется к временным последовательностям видео-, аудио- и графических данных. Действительно, любые данные, которые могут быть преобразованы в линейную последовательность, могут быть проанализированы с помощью DTW.

DTW заключается в измерении сходства между двумя последовательностями, которые могут меняться во времени или скорости. Для двух временных последовательностей $Q = q_1, q_2, \ldots, q_n$ и $C = c_1, c_2, \ldots, c_m$ это просто сумма квадратов расстояний от каждой k-ой точки одной последовательности до соответствующей точки другой. Расстояние DTW между двумя временными рядами рассчитывается на основе этого оптимального пути деформации, используя следующее уравнение:

$$DWT(Q, C) = min\left\{\frac{\sum_{k=1}^{K} d(w_k)}{K}\right\},$$

где $d(w_k) = d(q_i, c_j) = (q_i - c_j)^2$ – матрица расстояний [5–6].

K в знаменателе используется для нормализации различных путей деформации с разной длиной. Поскольку DTW должен потенциально исследовать каждую ячейку в матрице деформирования, ее пространственная и временная сложность – O(nm)

Hidden Markov Model

На смену алгоритма DTW пришел более совершенный подход – скрытые Марковские модели (HMM – Hidden Markov Model). HMM являются статистическими моделями, которые выводят последовательность символов или величин и используются для распознавания речи, поскольку речевой сигнал можно рассматривать как кусочно-стационарный сигнал или кратковременный стационарный сигнал [6–7]. HMM определяется как совокупность $\lambda = (A, B, \pi)$, где A – матрица вероятностей переходов, состоящая из элементов a_{ij} – вероятностей перехода из состояний i в j, B – матрица вероятностей наблюдения выходных значений, состоящая из элементов $b_i(o_k)$ – вероятностей наблюдения в состоянии j вектора признаков o_k, π – вектор вероятностей начальных состояний, состоящий из компонентов π_i – вероятностей нахождения в i-ом состоянии в начальный момент времени. Находясь в состоянии j в момент времени t, функция прямого распространения вероятностей определяется как вероятность наблюдения последовательности $O = (o_1, o_2, \ldots, o_t)$ [6]

$$\alpha_1(j) = \pi_j b_j(o_1),$$

$$\alpha_t(j) = \left[\sum_{i=1}^{N_s} \alpha_{t-1}(i) a_{ij}\right] b_j(o_t).$$

Вычисление $\alpha_t(j)$ происходит рекурсивно. Дойдя до конца наблюдаемой по-

следовательности, $\alpha_T(j)$ складывается для всех состояний, получив вероятность наблюдения исходной последовательности $O = (o_1, o_2, \ldots, o_T)$ [6],

$$P(O|\lambda) = \sum_{j=1}^{N_s} \alpha_T(j).$$

Данная вероятность используется при распознавании изолированных слов:

$$W^* = \arg\max P(O|\lambda).$$

Каждое слово или фонема имеет различное распределение выходных данных. Фонемы моделируются с использованием трех различных состояний – начального, среднего и конечного. Существует два типа фонем: монофоны и трифоны. У монофонов наложение артикуляции игнорируется, собираются модели фонем, стоящих отдельно. У трифонов наложение артикуляции учитывается, при этом происходит построение отдельной модели для фонем, окруженных другими фонемами. Скрытая марковская модель для ряда слов или фонем создается путем объединения отдельных скрытых марковских моделей для каждого слова или фонемы [1, 7].

Artificial Neural Networks

Для оптимизации алгоритма HMM часто используют нейронные сети, которые предварительно обрабатывают речевой сигнала, например преобразование объектов или уменьшение размерности. Искусственные нейронные сети (ANN – Artificial Neural Networks) – это вычислительные системы, основанные на биологических нейронных сетях, которые составляют мозг животных. Такие системы изучают (постепенно улучшают производительность) задачи, рассматривая примеры, как правило, без специального программирования. Нейронные сети представляют собой устройства для сопоставления образцов с архитектурой обработки, основанной на нейронной структуре человеческого мозга [2, 7]. Они состоят из простых взаимосвязанных блоков обработки (нейронов). Каждое соединение (синапс) между нейронами может передавать сигнал от одного к другому. Приемный (постсинаптический) нейрон может обрабатывать сигнал, а затем подключать к нему нейроны. В обычных реализациях ANN синапсовый сигнал является реальным числом, а выход каждого нейрона вычисляется нелинейной функцией суммы его входов [8].

$$p_j(t) = \sum_i o_i(t)\, w_{ij},$$

где w_{ij} – вес соединений.

Нейроны и синапсы обычно имеют вес, который корректируется по мере продолжения обучения. Вес увеличивает или уменьшает силу сигнала, который он посылает через синапс. Нейроны могут иметь такой порог, что только в том случае, если совокупный сигнал пересекает это пороговое значение, посылаемый сигнал.

Как правило, нейроны организованы в слои. Различные слои могут выполнять различные виды преобразований на своих входах. Сигналы перемещаются от первого (входного) к последнему (выходному) слою. При оценке вероятности сегмента речи нейронные сети позволяют проводить тестирование естественным и эффективным образом. Недостатком нейронных сетей является неспособность моделировать временные зависимости [1–2].

Deep Neural Network

Разновидностью нейронных сетей являются глубокие нейронные сети (DNN – Deep Neural Network). Данный алгоритм представляет собой искусственную нейронную сеть с несколькими скрытыми слоями единиц между входным и выходным уровнями. Подобно мелким нейронным сетям, DNN могут моделировать сложные нелинейные отношения. Архитектуры DNN создают композиционные модели, в которых дополнительные слои позволяют составлять элементы из нижних слоев, обеспечивая огромную учебную способность и, следовательно, потенциал моделирования сложных моделей речевых данных. DNN сеть имеет входной слой x, скрытый слой s и выходной слой y. Входной слой состоит из вектора $x(t)$, который является объединением вектора $w(t)$, представляющим собой текущее слово, и вектора $s(t-1)$, который представляет собой выходные значения скрытого слоя, полученные на предыдущем шаге. Размер вектора $w(t)$ равен размеру словаря. Выходной слой $y(t)$ имеет тот же размер, что и $w(t)$, и после изучения нейронной сети представляет собой вероятностное распределение следующего слова при данном предыдущем слове и состоянии скрытого слоя в предшествующий временной шаг [1–2]. Размер скрытого слоя обычно выбирается эмпирически. Все слои можно вычислить следующим образом:

$$s(t) = w(t) + s(t-1),$$

$$s_j(t) = f\left(\sum_i x_i(t) u_{ji}\right),$$

$$y_k(t) = g\left(\sum_j s_j(t) u_{kj}\right),$$

где $f(z)$ – сигмоидальная активационная функция:

$$f(z) = \frac{1}{1+e^{-z}},$$

$g(z)$ – функция softmax:

$$g(z_m) = \frac{e^{z}m}{\sum_k e^{z}k}.$$

Алгоритм End-To-End

На сегодняшний день наиболее современным алгоритмом является алгоритм End-to-End поиска вероятности возрастания, называемый LAS (Likelihood Ascent Search). LAS – это модель распознавания речи от конца до конца. LAS учится транскрибировать аудиопоследовательность сигнала к последовательности слов, по одному символу за раз, без использования явных языковых моделей, таких как HMM. Он состоит из энкодера, который называется listener, и декодера, который назван speller. LAS моделирует каждый выход символа y_i как условное распределение по сравнению с предыдущим символом [8–10]

$$p(y|x) = \prod_i p(y_i|x),$$

где $x = (x_1,\ldots, x_T)$ – входная последовательность, $y = (y_1,\ldots, y_s)$ – выходная последовательность, причем элемент множества y может являться любым значением из букв, цифр или знаков. Данная модель является дискриминирующей и сквозной, поскольку она непосредственно предсказывает условную вероятность последовательности символов, учитывая акустический сигнал [10].

Главным преимуществом систем распознавания речи стала дружественность к пользователю. Они позволяют вводить данные или команды посредством речи без использования сенсорных или иных методов [3]. Недостаток же заключается в неспособности распознавать некоторые вариации произношения, а также отсутствие поддержки большинства языков за пределами английского языка и невозможности сортировать фоновый шум. Такие факторы могут привести к неточностям [7, 9].

Распознавание речи имеет широкий спектр применения. Простые голосовые команды могут использоваться для инициирования телефонных звонков, выбора радиостанций или воспроизведения музыки с совместимого смартфона или MP3-плеера. Так же распознавание речи позволяет общаться на разных языках.

Система распознавания речи также используется в военных нуждах. Распознаватели речи успешно работают на военных самолетах с приложениями, включающими: настройку радиочастот, управление системой автопилота, настройку координатных частот и параметров выпуска оружия и контроль полета. За последние десятилетия на вертолетах были проведены значительные программы испытаний систем распознавания речи, в частности в рамках исследований и разработок авионики США (AVRADA – Aviation Research And Development Activity) и Королевского аэрокосмического учреждения (RAS – Royal Aeronautical Societ) в Великобритании. В ходе исследований была выявлена основная проблема – достижения высокой точности распознавания при шуме. Эта проблема является неразрешенной и по сей день.

Подводя итоги, следует отметить, что, хотя система распознавания речи уже развивается давно, ее нельзя назвать совершенной, поскольку она имеет ограниченный потенциал из-за своей тривиальности. Хотя автоматические системы распознавания речи далеко не идеальны с точки зрения точности слова или задачи, надлежащим образом разработанные приложения все еще могут эффективно использовать существующую технологию для предоставления реальной ценности клиенту, о чем свидетельствует количество таких систем, которые ежедневно используются миллионами пользователей. Для оптимизации распознавания речи необходимо иметь большую базу данных слов, произносимых разными людьми в различном эмоциональном состоянии, используя разные записывающие устройства (телефон, микрофон, прослушивающее устройство). На сегодняшний день развитие алгоритмов распознавания речи не прекращается. В дальнейшем можно прогнозировать развитие систем распознавания речи в области усовершенствования нейронных сетей. Также обязательным требованием станет наличие обратных связей на различных уровнях и разработкой новых методов обучения таких нейронных сетей.

Список литературы

1. Juang B.H., Lawrence R. Rabiner. Automatic Speech Recognition – A Brief History of the Technology Development, Georgia Institute of Technology, Atlanta, 2004 [Электронный ресурс]. – Режим доступа: http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf. (дата обращения: 17.12.2017).

2. Кипяткова И.С., Карпов А.А., Разновидности глубоких искусственных нейтронных сетей для систем распознавания речи // Труды СПИИРАН. – 2016. – № 6(49). – С. 80–103.

3. Белов Ю.С., Либеров Р.В. Подходы и проблемы распознавания личности по голосу // Электронный журнал: наука, техника и образование. – 2015. – № 3 (3). – С. 68–77.

4. Зулкарнеев М.Ю., Репалов С.А., Шамраев Н.Г. Система распознавания русской речи, использующая глубокие нейронные сети и преобразователи на основе конечных автоматов // Нейрокомпьютеры: разработка, применение. – 2013. – № 10. – С. 40–46.

5. Ghazi Al-Naymat, Sanjay Chawla, Javid Taheri, SparseDTW: A Novel Approach to Speedup Dynamic Time Warping, The University of New South Wales Sydney, 2012 [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1201.2969v1.pdf (дата обращения: 17.12.2017).

6. Огнев И.В., Парамонов П.А. Распознавание речи методами скрытых Марковских моделей в ассоциативной осцилляторной среде // Технические науки. Информатика, вычислительная техника. – 2013. – № 3(27). – С. 115–126.

7. Нифонтов С.В., Белов Ю.С. Применение скрытых марковских моделей в текстонезависимых системах идентификации пользователей по голосу // Электронный журнал: наука, техника и образование. – 2016. – № 2 (6). – С. 116–124.

8. Zell Andreas. «chapter 5.2». Simulation Neuronaler Netze [Simulation of Neural Networks], Addison-Wesley. – 1994.

9. Белов Ю.С., Нифонтов С.В., Азаренко К.А. Применение Вейвлет-фильтрации для шумоподавления в речевых сигналах // Фундаментальные исследования. – 2017. – № 4–1. – С. 29–33.

10. William Song, Jim Cai. End-to-End Deep Neural Network for Automatic Speech Recognition [Электронный ресурс]. – Режим доступа: http://cs224d.stanford.edu/reports/SongWilliam.pdf (дата обращения: 17.12.2017).