Метод повышения скорости работы декодера в задаче распознавания речи

Бинеев О.Р.
Зулкарнеев М.Ю.
Салман С.Х.

ФГНУ НИИ «Спецвузавтоматика»
Россия, г. Ростов-на-Дону, Газетный пер., 51.
Тел. (863) 297-50-84, факс (863) 297-50-84, asni@rnd.runnet.ru

Современные системы автоматического распознавания речи, основанные на скрытых марковских моделях, представляют собой сложные многопараметрические программные комплексы (особенно системы с большим словарём, где количество слов превышает 10^5), которые требуют тонкой многоэтапной настройки (обучения) и предъявляют высокие требования к используемой компьютерной технике как с точки зрения быстродействия, так и с точки зрения используемой памяти. Несмотря на то, что в настоящее время разработаны эффективные алгоритмы декодирования, добиться работы декодера в реальном масштабе времени с сохранением высокого уровня точности по-прежнему сложно. В этой работе предлагается подход к ускорению работы динамического однопроходного Витерби-подобного декодера с древовидной структурой сети распознавания, который используется при распознавании речи с большим словарём. Основная вычислительная нагрузка при работе декодера приходится на вычисление отклика гауссовых смесей, моделирующих состояния контекстнозависимых фонем. В работе при вычислении откликов предлагается использовать алгоритм «дорожная карта», который позволяет находить l лучших гауссоид (дающих наибольший отклик) для данного наблюдения без вычисления откликов всех гауссоид. Перед выполнением декодирования для каждой гауссоиды находится список наиболее близких гауссоид с использованием в качестве расстояния перекрытия данных гауссоид в пространстве признаков. При декодировании выполняется поиск гауссоид, дающих наилучший отклик для данного наблюдения. Процедура поиска является итерационной и напоминает прокладывание маршрута по карте (отсюда название алгоритма).

ВВЕДЕНИЕ

Технология автоматического распознавания речи, основанная на скрытых марковских моделях (СММ) и n-граммных моделях языка [1], в настоящее время является наиболее популярной при создании систем распознавания речи. С развитием компьютерной техники повышается сложность систем, основанных на этой технологии. Так, если в 70–80-х годах XX века такие системы были способны распознавать раздельные слова со словарём размером 100–1000 слов, то в 90-х годах появились системы распознавания непрерывной речи с размером словаря в десятки тысяч слов.

Сейчас на повестке дня стоит задача создания системы распознавания речи с размером словаря, превышающем 10^6. Ограничения на увеличение раз-

мера словаря слов устанавливает главным образом декодер. Существуют различные типы декодеров, используемых в системах распознавания речи [2]. В этой работе используется декодер, основанный на алгоритме перемещающегося маркера [3], который является практической реализацией алгоритма Витерби [2]. В нём в качестве оптимального частичного пути используется объект, который называется «маркер», при этом переходы между состояниями заданы явно посредством сети распознавания (рис. ***1***).

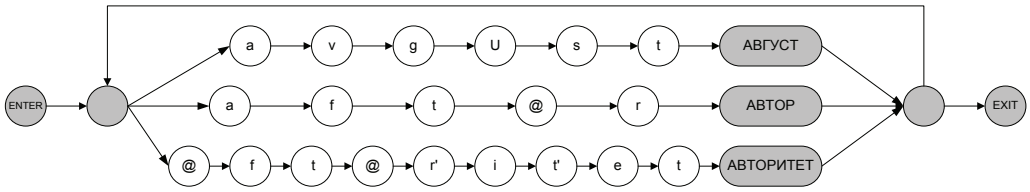

Рис. 1. Пример сети распознавания

Эксперименты показывают [4], что декодер, использующий такую сеть распознавания, со словарем размером больше несколько тысяч. А использование трёхграммной модели языка для такого декодера и вовсе невозможно. В работе [4] используется модифицированная сеть распознвания. В ней одинаковые начальные части фонетических транскрипций различных слов объединены. Пример сжатой сети приведён на рисунке 2. В ней начальная фонема «а» слов «АВГУСТ» и «АВТОР» представлена одним и тем же узлом сети.

Использование сжатой сети распознавания позволяет значительно увеличить скорость декодирования по двум причинам. Во-первых, с уменьшением количества узлов уменьшается количество маркеров. Во-вторых, количество маркеров зависит от номера фонемы в слове. Поскольку для такой сети количество узлов, соответствующих начальным фонемам, гораздо меньше количества узлов конечных фонем, количество маркеров снижается ещё.

В данной работе для ускорения работы декодера предлагается использовать алгоритм «Дорожная карта» [5]. Он позволяет находить наиболее вероятные компоненты гауссовых смесей, без необходимости рассчитывать их все. В следующем разделе даётся более подробное описание метода, а далее приводятся результаты его экспериментальной проверки.

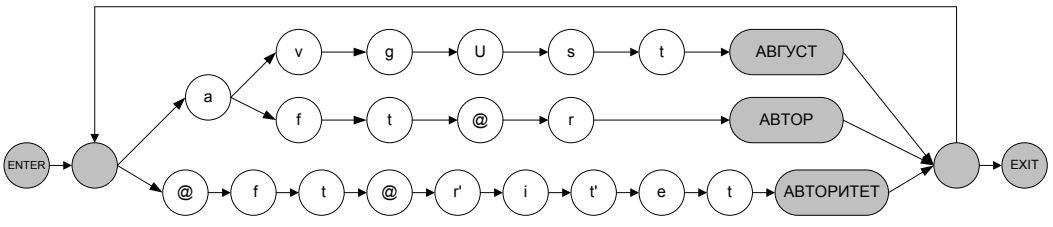

Рис. 2. Сжатая сеть распознавания

ОПИСАНИЕ МЕТОДА

В работе предлагается метод ускорения работы декодера за счёт уменьшения количества вычислений. Далее описание алгоритма «Дорожная карта» ведётся в соответствии с работой [5].

На каждом шаге декодирования требуется вычисление выходной вероятности $b_j(ot)$ для всех состояний, в которые есть переходы из состояний, содержащих маркеры. Вычисление

$b_j(ot)$ является наиболее ресурсоёмкой частью процедуры декодирования, поскольку для вычисления смеси $b_j(o)=\sum_{m=1}^{M} w_{jm} N(o|\mu_{jm}, \Sigma_{jm})$ требуется вычисление всех её компонент. В работе предлагается не рассчитывать все компоненты всех смесей, имеющихся в системе, а найти l наиболее вероятных компонент, а при вычислении $b_j(ot)$ использовать аппроксимацию $b_j(o) \approx N(o|\mu_{jm}, \Sigma_{jm})$, если *l*-я компонента попала в найденный список, и $b_j(o)=0$, если ни одна из компонент смеси, описывающей $b_j(o)$, в список не попала. Для нахождения l наиболее вероятных компонент предлагается использовать алгоритм «дорожная карта», который позволяет находить l лучших компонент (дающих наибольший отклик) для данного наблюдения без вычисления откликов всех компонент. Пусть M — множество всех компонент и для каждой компоненты $m \in M$ известен список ближайших к ней компонент $n(m)$. Дорожная карта — это граф связей компонент друг с другом, которые задаются списками $n(m)$.

Алгоритм «Дорожная карта» является итерационным и состоит из следующих шагов:

1. Инициализация результирующего списка ℓ.
2. Выбор l наиболее вероятных компонент из множества $\ell \cup n(\ell)$ в качестве нового списка $\bar{\ell}$.
3. Если $\bar{\ell}=\ell$, выбор l наиболее вероятных компонент из множества $\ell \cup r(M)$ в качестве нового списка ℓ, иначе возвращение к шагу 2.
4. Выход, если $\bar{\ell}=\ell$, иначе возвращение к шагу 2.

Начальный список может быть задан случайно или в качестве начального списка может быть взят список с предыдущего шага декодирования. $r(M)$ — случайным образом выбранное подмножество множества M.

Инициализация списков $r(M)$ наиболее близких к компоненту m компонент выполняется похожим образом, только в этом случае в качестве наблюдения o выступает компонента m:

1. Инициализация списков $n(m)$ для всех $m \in M$.
2. Выбор в качестве нового списка $\bar{n}(m)$ наиболее близких к компоненте m компонент из множества $n(m) \cup n(n(m))$ для всех $m \in M$.
3. Если $\bar{n}(m)=n(m)$, выбор n наиболее близких к компоненте m компонент из множества $n(m) \cup r(M)$ в качестве нового списка $\bar{n}(m)$, иначе возвращение к шагу 2.
4. Выход, если $\bar{n}(m)=n(m)$, иначе возвращение к шагу 2.

В качестве расстояния между двумя компонентами $m^{(1)}$ и $m^{(2)}$ используется их перекрытие в пространстве признаков $\delta(m^{(1)}, m^{(2)})$, для вычисления которого используются выражения:

$$\delta(m^{(1)}, m^{(2)}) = \sum_{i=1}^{d} -\log(O(m_i^{(1)}, m_i^{(2)}))$$

$$O(m_i^{(1)}, m_i^{(2)}) = \int \min(O(m_i^{(1)}(O), m_i^{(2)}(o)))\, do \qquad (1),$$

где $O(m_i^{(1)}, m_i^{(2)})$ — перекрытие двух одномерных нормальных распределений $N(\mu_i^{(1)}, \Sigma_{ii}^{(1)})$ и $N(\mu_i^{(2)}, \Sigma_{ii}^{(2)})$ (см. рисунок 3), $\mu_i^{(1)}, \Sigma_{ii}^{(1)}, \mu_i^{(2)}, \Sigma_{ii}^{(2)}$ — компоненты векторов средних и ковариационных матриц многомерных нормальных распределений $N(\mu^{(1)}, \Sigma^{(1)})$ и $N(\mu^{(2)}, \Sigma^{(2)})$, соответствующих компонентам смесей $m^{(1)}$ и $m^{(2)}$ соответственно.

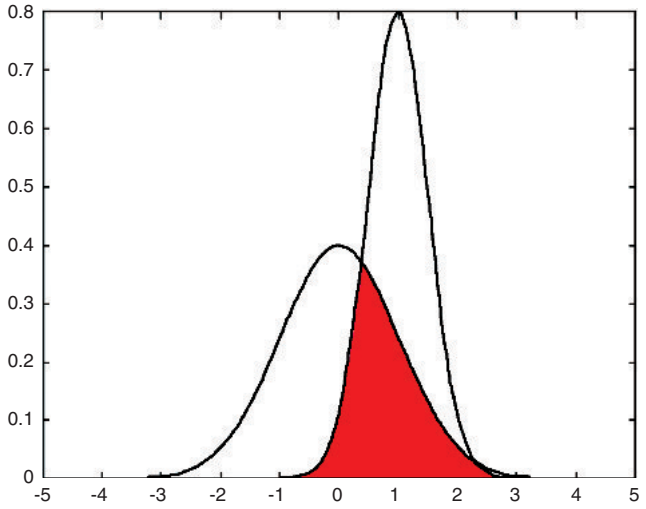

Рис. 3. Перекрытие двух одномерных нормальных распределений

Для вычисления логарифма интеграла (1) используется аппроксимация

$$\log(O(m_i^{(1)}, m_i^{(2)})) \approx$$

$$\approx \frac{-0.0557}{\sigma} + 0.3137\sigma - 0.3292\mu^2 - 0.0037\frac{\mu^2}{\sigma} + 0.0429\mu^2\sigma - (0.3137 - 0.0557)$$

где $\sigma = \frac{\sigma_2}{\sigma_1}$, $\mu = \frac{\mu_2 - \mu_1}{\sigma_1}$. Здесь предполагается, что $\sigma_2 \le \sigma_2$, в противном случае компоненты меняются местами.

ОПИСАНИЕ ЭКСПЕРИМЕНТОВ

Для проверки предлагаемого метода были проведены эксперименты по полнотекстовому распознаванию с использованием микрофонной речевой базы русского языка. Речевая база была разбита на две части. Первая часть длительностью 20 часов была использована для обучения трифонных моделей фонем. В результате обучения было получено 16 тыс. связанных трифонных моделей с общим количеством различных компонент смесей около 4000 тыс. В качестве языковой модели использовалась трёхграммная модель языка. Для декодирования использовался декодер с сетью распознавания, пример которой приведен на рисунке 1. Вторая часть речевой базы длительностью 1 час использовалась для тестирования. Все слова из тестирующей выборки содержались в словаре распознавания.

Было проведено два эксперимента: 1) эксперимент, использующий стандартный декодер; 2) эксперимент использующий алгоритм «Дорожная карта».

Результаты экспериментов приведены в таблице. В первом столбце приводится точность распознавания слов $\frac{N_h}{N}$, где N_h — количество правильно распознанных слов, N – общее количество слов в тестирующей выборке. Во втором столбце приводится время распознавания, нормированное на длительность тестирующей выборки.

Таблица. Результаты экспериментов

Эксперимент	Точность распознавания слов, N_h/N, %	Время распознавания, RT
Стандартный декодер	74,1	4,7
«Дорожная карта»	73,6	3,2

ЗАКЛЮЧЕНИЕ

Результаты экспериментов показали, что использование алгоритма «Дорожная карта» позволяет увеличить быстродействие в 1,5 раза. Кроме этого, снижение количества откликов, которые надо рассчитать, позволяет надеяться на дальнейшее увеличение скорости обработки. Для этого в дальнейшем планируется выполнить дополнительную оптимизацию алгоритма.

ЛИТЕРАТУРА

1. *Рабинер Л*. Скрытые Марковские модели и их применение в избранных приложениях при распознавании речи: Обзор 2, февраль 1989 г., ТИИЭР, Т. 77, стр. 86-120.

2. *Xuedong Huang, Alex Acero and Hsiao-Wuen Hon* Spoken Language Processing, A Guide to Theory, Algorithm and System Development. New Jersey : Prentice Hall Inc., 2001.

3. *Young S. J.* Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems. 1989 : s.n., CUED Technical Report F INFENG/TR38 Cambridge University.

4. *Odell, J. J. et al*. A One Pass Decoder Design for Large Vocabulary Recognition. 1994. Proceedings ARPA Workshop on Human Language Technology. pp. 405-410. Merrill Lynch Conference Centre.

5. *Povey D. and Woodland P.C*. Frame discrimination training of HMMs for large vocabulary speech recognition. Cambridge university engineering department. Cambridge : s.n., 2000. Technical report.