А Н А Л И З Т Е К С Т О В И С И Г Н А Л О В
(Вычислительные системы)
1987 год Выпуск 123

УДК 621.391

ОБУЧЕНИЕ В РАСПОЗНАВАНИИ СЛИТНОЙ РЕЧИ

В.М.Величко

Рассматриваются два вопроса, касающиеся обучения в распознавании слитной речи: выбор элементарных единиц представления речевого сигнала и выделение эталонов слов и морфем из слитных словосочетаний.

Для выбора элементарных единиц представления речевого сигнала применен метод таксономии [I] в пространстве первичного описания, называемый в литературе по речевым исследованиям методом векторного квантования [2].

Для выделения эталонов слов или морфем из слитных словосочетаний применяется алгоритм динамического программирования [4] с определением оптимального пути и автоматическим выделением нужного участка из слитного сочетания.

I. Применение таксономии для выбора представления речевого сигнала

При построении систем распознавания и понимания слитной речи на базе микрокомпьютеров существенные трудности возникают из-за значительного увеличения объема вычислений по сравнению с задачей распознавания ограниченного набора изолированных команд. Это вызвано большим перебором вариантов при многообразии допустимых сочетаний слов во фразах.

Одним из способов ускорения принятия решений является переход от векторного представления коротких сегментов речевого сигнала к символьному, например к фонемной маркировке сегментов. В последние годы получило распространение векторное квантование как форма представления сегментов в виде номеров векторов, хранящихся в так называемой кодовой книге объемом от нескольких десятков до

нескольких тысяч векторов. Векторы выбираются по определенным правилам на основе обучающей выборки и составляют алфавит для отображения последовательности сегментов обучающей выборки и контрольной (т.е. подлежащей распознаванию) реализации в последовательность символов.

Очевидным преимуществом подобного представления являются следующие возможности:

- значительное сокращение памяти, требуемой для хранения эталонов;

- предварительное вычисление расстояний между векторами в любой сколь угодно сложной метрике и хранение вычисленных расстояний в таблице с последующим извлечением их из памяти без дополнительных вычислительных затрат;

- применение текстовых методов обработки контрольной реализации (т.е. методов, основанных на учете наличия или отсутствия определенных таксонов в контрольной реализации) для быстрой предварительной отбраковки бесперспективных эталонов (см., например, [5]).

Идея применения метода таксономии для распознавания речи была выдвинута и частично проверена в 1967 г. и позже в работах коллектива Института математики СО АН СССР под рук. проф. Н.Г. Загоруйко (в частности, в первых экспериментах работы [4]). В последнее время эта идея получила широкое распространение и под названием метода векторного квантования применяется не только в распознавании речи, но и в синтезе, а также передаче речевых сообщений по каналам связи. Смысл метода состоит в разбиении по какому-то критерию пространства признаков речевого сигнала на участки-таксоны в зависимости от наличия в них реальных представителей речевого сигнала обучающей выборки. Каждому участку присваивается номер, и в дальнейшем каждый сегмент речевого сигнала заменяется номером таксона, в который попадает этот сегмент.

В данной работе проверялись различные приемы разбиения пространства признаков на таксоны и оценивалась их эффективность по результатам распознавания конкретных словарей.

В работах [1,2] описаны различные алгоритмы таксономии, дающие приемлемые результаты. Однако, учитывая специфику реализации алгоритмов обучения на микрокомпьютере, мы использовали упрощенные приемы, использующие небольшой объем памяти и вычислений при удовлетворяющем пользователя качестве разбиения.

При равномерном разбиении пространства признаков распределение эталонных элементов по таксонам будет крайне неравномерным - подавляющая часть объема признакового пространства останется пустой, а в некоторых частях будут сосредоточены разнородные элементы. Такое разбиение эквивалентно скалярному квантованию. Поэтому требование зависимости разбиения от реальных эталонных сегментов является существенным.

Проверяемые алгоритмы строились следующим образом. Задана обучающая выборка речевого сигнала в виде последовательности N-мерных векторов, соответствующей однократному произнесению каждого из слов распознаваемого словаря. В описываемом эксперименте N = 6, длительность отрезка речевого сигнала, соответствующего вектору, составляет 16 мсек. Задаются ограничения на общее число таксонов (длину кодовой книги) и на пороговое число R - радиус таксона. Программа таксономии выбирает один из сегментов обучающей выборки $\vec{X}(x_1, \ldots, x_N)$ и полным перебором определяет все сегменты $\vec{Y}$, удаленные от выбранного не более чем на пороговое расстояние, т.е. удовлетворяющие условию:

$$D(\vec{X},\vec{Y}) = \sum_{i=1}^{N} |x_i - y_i| \leq R. \qquad (*)$$

При этом для сокращения времени во избежание повторной проверки программа помечает отобранные сегменты. Множество таких сегментов образует очередной таксон. Затем выбирается следующий неиспользованный сегмент, и процедура повторяется до исчерпания обучающей выборки либо по достижении предельного числа таксонов (в этом случае надо принимать специальные меры). На выходе программы формируются перечень таксонов с координатами соответствующих векторов и обучающая выборка в виде последовательности номеров таксонов вместо последовательности векторов.

Проверяемые алгоритмы таксономии отличались друг от друга:

- способом выбора начального и последующих исходных векторов;
- способом задания вектора, определяющего координаты таксона;
- критерием окончания таксономии;
- наличием или отсутствием ограничений снизу на количество элементов в таксоне.

Наши эксперименты 1967 г. в рамках работы [4] дали невысокую надежность (~90%), как оказалось, из-за малого (до 37) количества использованных таксонов.

Стимулом для возобновления экспериментов послужили весьма высокие результаты распознавания речи, описанные в работе [3]. Анализ наших результатов распознавания конкретных словарей с использованием таксономии привел к выводу о том, что качество распознавания повышается с уменьшением радиуса таксона, т.е. с повышением точности представления сегмента, и, как следствие, с увеличением числа таксонов. Поэтому основные усилия были направлены на создание простых алгоритмов, дающих минимальный радиус таксона при фиксированном допустимом числе таксонов. Предельное число таксонов было выбрано равным 255, чтобы номер таксона мог быть представлен одним байтом. Это достаточно хорошо соответствует и представлениям фонетистов о разнообразии аллофонов русского языка (200-300).

Известно [1], что для эвристических алгоритмов таксономии порядок перебора элементов оказывает существенное влияние на результат. Первоначально в качестве исходного элемента брался сегмент, наиболее удаленный от центра обучающей выборки, затем наиболее удаленный от предыдущего исходного элемента и т.д. Казалось, что такая схема наилучшим образом учтет удаленные сегменты и приведет к минимальному количеству таксонов, представленных единственным сегментом. Однако среднее число сегментов в таксоне при этом оказалось небольшим, а радиус при предельном числе таксонов был достаточно велик.

Метод простого последовательного просмотра сегментов, начиная с первого, оказался вполне приемлемым по количеству и размеру таксонов. Он имеет также дополнительное преимущество - возможность осуществления таксономии по мере поступления эталонов, что существенно при большом объеме словаря, когда вся выборка в векторном виде не помещается в оперативную память.

По способу задания вектора, определяющего координаты таксона, сравнивались прямое использование исходного вектора в качестве координат таксона и усредненное значение векторов сегментов, попавших в таксон. Разница оказалась довольно значительной в пользу усреднения: 16 ошибок против 24 на одной и той же обучающей и контрольных выборках. Казалось бы, исходный вектор, формирующий таксон, и должен находиться в центре таксона. Однако разница между

исходным вектором и центром таксона при усреднении фактически составляла от 42% до 77% (в среднем 55%) от радиуса таксона. Объяснение этого неожиданного преимущества усреднения состоит в том, что в многомерном пространстве основной объем шара приходится на тонкий слой вблизи его поверхности, где преимущественно располагаются точки, составляющие таксон (например, в 6-мерном пространстве половина объема шара сосредоточена в поверхностном слое толщиной в 11% от радиуса шара). Усреднение переводит центр таксона туда, где вероятность нахождения вектора реального сигнала очень мала, а выигрыш по точности представления таксона оказывается значительным.

Критерием окончания таксономии, очевидно, является момент исчерпания всех сегментов обучающей выборки. Однако при небольшом пороге таксономии число таксонов может достичь предельной величины (255). Был проверен вариант неполной таксономии, т.е. разбиения до предельного количества таксонов, а затем оставшимся сегментам обучающей выборки присваивался номер ближайшего таксона. Этот прием позволил несколько уменьшить радиус таксона и соответственно улучшить качество распознавания. В табл. I приведены для сравнения результаты распознавания в зависимости от порога и числа сегментов, попавших в таксоны при полной и неполной таксономии, на словаре объемом в 129 слов [6] для одной и той же контрольной выборки.

Таблица I

Порог R	Число таксонов	Сегменты, %	Ошибки, %
90	255	64	3,1
100	255	78	3,1
110	255	98	3,9
120	216	100	4,65
130	166	100	5,4

Однако при неполной таксономии нецелесообразно ограничивать количество таксонов путем отбрасывания конца обучающей выборки, так как в нем могут оказаться самостоятельные крупные таксоны. Поэтому было проверено еще одно условие, а именно ограничение на минимальное число сегментов, содержащихся в таксоне. При проведении таксономии подсчитывалось число сегментов в каждом таксоне (это необходимо и для усреднения), и при числе сегментов, меньшем заданного порога, таксон аннулировался, а его сегментам присваивались номера ближайших таксонов. Такой прием позволил значительно сократить радиус сформированных таксонов при соответствующем повышении качества распознавания.

Таблица 2

Порог R	Число таксонов	Сегменты, %	Ошибки, %
96	255	100	12,7
80	201	90,4	12,4
70	237	84	12,4
64	250	79	8,0
60	244	72	7,8
Б/Т	-	-	11,2

Итоговые результаты распознавания приведены в табл.2. Проверка проводилась на очень трудном для распознавания словаре из 67 слов, содержащем группы слов, отличающихся одной фонемой (колос - голос - полоз - волос, дата - та - та - вата - хата и т.д.). Порог по минимальному числу сегментов в таксоне везде был равен 3, кроме первого эксперимента (строка 1), где он был равен 1. В последней строке для сравнения приведены результаты распознавания без применения таксономии для той же обучающей выборки (контрольные выборки при этом были другие).

Во всех экспериментах использовался один и тот же алгоритм распознавания, основанный на локально-оптимальном методе динамического программирования (ДП), с адаптивным коридором и с отсечками бесперспективных эталонов по длительности, интегральным характеристикам слов, а также глобальным и локальным критериям оценки расстояний между реализациями слов [6]. При распознавании расстояние D между сегментами эталонов T и контрольной реализацией $\vec{X}$ определялось по формуле (*):

$$D(\vec{X},\vec{T}) = \sum_{i=1}^{N} |x_i - t_i| .$$

Оценим выигрыш в памяти, полученный благодаря применению метода таксономии.

Для хранения эталонов K слов обучающей выборки в векторном представлении и таблицы, содержащей информацию об адресах и длинах слов-эталонов в массиве, требуется $K \cdot N \cdot L + 2 \cdot 2 \cdot K$ байт, где N - размерность признакового пространства, L - средняя длина слова в сегментах. Предполагается, что векторы сегментов требуют по одному байту на компонент, а информация - по 2 байта на число.

Для хранения эталонов K слов обучающей выборки в виде номеров таксонов и той же информационной таблицы требуется

$$K \times L + 2 \times 2 \times K + 2 \times N \times 255 \text{ байт.}$$

Таблица 3

Объем словаря (К слов)	Память при векторном представлении (Кбайт)	Память при таксономии (Кбайт)
100	21,5	7
200	43	11
500	107	22,5

Последнее слагаемое отражает объем памяти для хранения по отдельности координат таксонов признакового пространства сегментов и интегральных характеристик слов, предназначенных для предварительной отсечки бесперспективных эталонов [6].

Приведем оценки памяти для словарей различного объема (табл. 3).

Приведенные результаты свидетельствуют о целесообразности применения таксономии при распознавании речи с точки зрения получения достаточно высокой надежности и экономичного представления по памяти.

2. Применение метода динамического программирования для выделения эталонов слов и морфем из слитных словосочетаний

Другим резервом повышения качества распознавания слитной речи является формирование эталонов с учетом слитного произнесения.

Обычной методикой обучения при распознавании фраз слитной речи, составленных из слов заданного словаря, является произнесение изолированных слов этого словаря и использование их в качестве эталонов при распознавании. Однако известно, что при слитном произнесении наблюдается эффект коартикуляции, заключающийся во взаимном влиянии соседних звуков с изменением их фонетического качества. Это проявляется, например, в оглушении или озвончении конечных согласных. Кроме того, меняется характер ударения внутри фразы, что также приводит к изменению фонетического качества звука.

Чтобы полнее учесть вариацию контрольной реализации по сравнению с эталонной, предлагается в рамках традиционной схемы обучения добавить эталоны слов, образованные при естественном слитном произнесении. Для выделения эталонов можно применить метод ДП, используя изолированные эталоны в качестве составляющих эталона произнесенной слитной фразы и находя соответствие между словами контрольной и эталонной последовательностей. Тот же подход может быть применен для выделения эталонов морфем из слов и слитных словосочетаний.

Эксперименты по проверке методики выделения эталонов прово - дились на материале слитно произносимых чисел от нуля до 999.Словарь включает в себя 37 слов, которые можно разбить на пять групп: I-я группа - I-9 (единицы); 2-я группа - I0-I9; 3-я группа - 0 (ноль); 4-я группа - 20-90 (десятки); 5-я группа - I00-900 (сот - ни).

Разбиение на группы сделано для учета семантических ограни - чений при распознавании фраз, представляющих числа в требуемом диапазоне. Эти ограничения существенно использовались при про - верке эффективности описываемого метода обучения для распознава - ния слитной речи. Они состоят в следующем: число может начинаться и заканчиваться словами любой группы; за словами I-3-й групп может следовать только конец фразы; за словами 4-й группы - I-я группа; за словами 5-й группы - I,2 и 4-я группы.

Выбор эталонов, подлежащих дублированию в слитных словосочетаниях, производился по результатам предварительного распознавания с использованием только изолированно произнесенных эталонов. При этом обнаружились плохо распознаваемые слова, которые,в основном, проявлялись на стыке фонетически подобных участков (например, 22, II8, 9I) или при изменении качества звуков (например, шипящих - 76, 560, 27, безударных - I20).

В процессе обучения система запрашивает пользователя, какое слово и из какого словосочетания необходимо записать в качестве эталона (например, 20 в сочетании 22). Затем система составляет эталонную фразу, состоящую из эталонов изолированно произнесенных заданных слов. Методом динамического программирования (ДП) опре - деляется оптимальный путь, который позволяет найти границы требуемого дополнительного эталона. В методе ДП использовались следующие рекуррентные соотношения:

$$S_{ij} = \min(S_{i-1,j}, S_{i,j-1}, S_{i-1,j-1}) + D(\vec{X}_i, \vec{Y}_j),$$

здесь S_{ij} - длина пути в точке i,j матрицы расстояний; D - расстояние по формуле (*).

При запоминании матрицы оптимальных путей с целью экономии памяти использовалось по 2 бита для каждого из трех возможных локально-оптимальных путей: слева, сверху и по диагонали слева-сверху направо-вниз. Кроме того, составляющие матрицы определялись только в коридоре вдоль диагонали шириной I4.

Для большей точности выделения эталона процедура, как правило, повторялась с использованием вновь полученного эталона в качестве составной части эталонного словосочетания. Общее число эталонов может колебаться в зависимости от качества распознавания. В наших экспериментах оно было принято равным 60 для 37 слов числового словаря. При этом представителей I-й группы было I5, 2-й - I3; 3-й - I; 4-й - I6; 5-й - I5.

Исходная пословная надежность распознавания с применением только изолированных эталонов была довольно низкой и равнялась 89,8% (43 ошибки при произнесении 209 чисел, содержащих 421 слово). После введения добавочных эталонов надежность повысилась до 94,5% (25 ошибок для 209 чисел с 453 словами).

Та же методика была применена при выделении квазиморфем из чисел вышеописанного словаря. При этом ожидаемым преимуществом, кроме некоторой экономии памяти, было повышение надежности распознавания за счет одинаковых частей похожих слов. Как известно, ошибки распознавания при использовании метода ДП возникают из-за того, что случайный шум в оценке меры расстояния на одинаковых словах может превысить разницу в расстояниях на отличающихся участках разных слов. При использовании одинаковых эталонов квазиморфем в разных словах они дадут в точности одинаковый шум, и вся разница в расстояниях будет образована только за счет фонетически различающихся участков слов.

В качестве квазиморфем применялись следующие единицы (в скобках указаны числительные, в которых используется данная единица): ОДИ'Н (I,II), ДВА' (2,20), ДВЕ (I2), ТРИ' (3,30,300) ТРИ (I3), ЧЕТЫ'РЕ (4,400), ЧЕТЫ'Р (I4), ПЯ'ТЬ (5), ПЯТ (I5,500), ПЯТЬ (50), ШЕ'СТЬ (6), ШЕС (I6,600), ШЕ СЬ (60), СЕ'МЬ (7,70), СЕМ (I7,700), ВОСЕ'МЬ (8,80), ВОСЕМ (I8,800). ДЕ'ВЯТЬ (9), ДЕВЯТ (I9,900), ДЕВЯНО' (90), ДЕ'СЯТЬ (I0), ДЕСЯ'Т (50,60), ДЕСЯТ (70,80), НАДЦАТЬ (II,I4), НА'ДЦАТЬ (I2, I3, I5-I9), НО'ЛЬ (0), ДЦАТЬ (20,30), СО'РОК (40), СТО' (I00), СТА (90,300,400), СО'Т (500,900), ДВЕ'СТИ (200), - всего 32 квазиморфемы.

Проверка распознавания проводилась на словаре изолированно произнесенных числительных. В качестве эталонов использовались последовательности составляющих слово квазиморфем. В результате испытаний обнаружилось, что практически исчезли обычные для этого словаря ошибки типа I2-I3-I5-I9. Результаты распознавания близки к I00%. Впрочем, и при обычных методах надежность распознавания этого словаря достаточно высокая. Преимущества квазиморфемно-

го подхода должны проявиться при распознавании слитной речи. Но объем проведенных экспериментов пока недостаточен для количественного подтверждения этого предположения.

3. З а к л ю ч е н и е

Проведенные эксперименты подтвердили целесообразность использования при обучении и распознавании слитной речи методов таксономии и ДП для формирования более экономичной и качественной обучающей выборки. Дальнейшие исследования направлены на развитие текстовых методов обработки информации при распознавании слитной речи для повышения быстродействия и надежности системы, а также на создание удобного сервиса для обучения пользователя.

Л и т е р а т у р а

I. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение. М.: Сов.радио, 1972. - С.87-116.

2. Векторное квантование при кодировании речи //Маккоул Дж., Рукос С., Гиш Г., ТИИЭР, пер. с англ., 1985. - Т.73. - № 11.-С.19-61.

3. ДЖЕЛИНЕК Ф. Разработка экспериментального устройства,распознающего раздельно произносимые слова //Там же. - С. 91-100.

4. ВЕЛИЧКО В.М., ЗАГОРУЙКО Н.Г. Автоматическое распознавание ограниченного набора устных команд //Вычислительные системы.- Новосибирск, 1969. - Вып. 36. - С. 101-110.

5. ВЕЛИЧКО В.М., ЗАГОРУЙКО Н.Г. Распознавание больших словарей //Автоматическое распознавание слуховых образов (АРСО-УШ).Ч.3. Тез. докл. УШ Всесоюзного семинара 16-23 сент. 1974 г.: Львов, 1974. - С.9-13.

6. ВЕЛИЧКО В.М. Минимизация вычислений в распознавании речи //Анализ символьных последовательностей.-Новосибирск, 1985. -Вып. 113. - Вычислительные системы. - С. 123-132.

7. ВЕЛИЧКО В.М. Алгоритм распознавания слитной речи с использованием семантико-синтаксических ограничений //Автоматическое распознавание слуховых образов/ Тезисы докл. и сообщ. 12-го Все-союз. семинара "Автоматическое распознавание слуховых образов" (АРСО-12), сент. 1982. Киев-Одесса, 1982. - С.342-345.

Поступила в ред.-изд.отд.
28 мая 1987 года