OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 004.89:004.4

МЕТОДЫ СЕМАНТИЧЕСКОГО АНАЛИЗА ДЛЯ ПОСТРОЕНИЯ ГОЛОСОВЫХ ИНТЕРФЕЙСОВ: РАСПОЗНАВАНИЕ РЕЧИ

И.Э.Хейдоров (*igorhmm@mail.ru*)
Белорусский государственный университет, Минск, РБ

В данной статье анализируется современная структура систем распознавания речи и предлагаются методы улучшения точности распознавания речи путем использования методов семантического анализа.

Введение

Компьютерная техника успешно развивается уже на протяжении полувека. За это время пройден колоссальный путь от примитивных вычислительных устройств размером в дом до супермощных портативных устройств, способных в реальном времен решать задачи огромной сложности. Ни одна другая научно-техническая отрасль не развивалась такими темпами и не достигла таких успехов, что во многом способствовало бурной информатизации всего общества.

Одновременно с появлением первых компьютеров пристальное внимание уделялось развитию средств ввода-вывода информации и созданию эффективных и удобных интерфейсов “человек-компьютер”. В этой связи успехи в области создания вычислительных средств мощный импульс получили исследования в области обработки, распознавания и синтеза речи как наиболее удобного и естественного для человека средства коммуникации [2, 5, 6, 9]. Создание программно-аппаратных средств для обеспечения речевых интерфейсов стало одним из приоритетных направлений научно-технического развития в области ИТ. За последние пятьдесят лет в эту сферу были вложены сотни миллиардов долларов в большинстве ведущих стран мира.

Однако несмотря на огромные научно-технические и финансовые ресурсы, вложенные в развитие этой отрасли, успехи в данном направлении нельзя признать абсолютными. Существующие в настоящий момент системы распознавания речи обладают большим количеством ограничений, точность их работы в значительной степени зависит от условий эксплуатации, начального обучения, правильности произношения слов целевым диктором и т.д., что сильно ограничивает сферу и надежность их использования. В области синтеза речи ситуация лучше, вполне приемлемые по качеству системы генерации речи существуют сейчас для большинства мировых языков. Однако для каждого языка существует своя специфика, которая не позволяет создать идеальную систему синтеза с использованием имеющихся подходов.

И самое главное, только сейчас исследователи вплотную приступили к решению основной задачи естественного интерфейса - пониманию речи и генерации соответствующего голосового ответа. Без решения данной задачи невозможно создание эффективного голосового интерфейса для использования в самых разнообразных областях науки и техники - от естественного интерфейса мобильных телефонов до полноценного управления роботом. Так в чем же основная проблема, что полный голосовой интерфейс еще не создан, несмотря на все титанические усилия?

Процедура распознавания речи включает в себя целый ряд этапов, требовательных к вычислительным ресурсам [1, 3, 4, 12, 17, 20]. Еще несколько десятилетий назад недостаток

мощности значительно ограничивал возможности применения ресурсоемких алгоритмов при построении систем распознавания речи. В настоящий момент таких ограничений практически не существует, однако существенного прорыва в области создания голосовых интерфейсов вычислительных устройств не произошло. Почему? Очевидно, что необходимо использовать новые знания и новые подходы для построения полноценных естественных интерфейсов “человек-компьютер”. В данной статье анализируются современные методы построения систем распознавания и предлагается новая схема построения голосовых интерфейсов.

1 Общая структура системы распознавания речи

Общая структура системы распознавания слитной речи (РСР) может быть разделена на следующие этапы [4]:

- предварительная обработка аудиосигнала и извлечение вектора признаков;
- сегментация аудиосигналов на участки речь-музыка-тишина-речь с фоновым шумом (для выделения фрагментов данных, содержащих только речь);
- выбор базовой акустической единицы речи;
- акустическое моделирование речевых сигналов;
- лингвистическое моделирование, позволяющее использовать априорные знания о структуре и особенностях языка;
- поиск всех возможных вариантов слов в слитной речи, оценка достоверности гипотез;
- верификация найденных слов.

Общая схема работы системы РСР представлена на рисунке 1.

Система РСР включает в себя четыре основных блока. Первый блок позволяет выделить речевые сегменты из слитного потока аудио на входе системы, и представить их в виде последовательности векторов признаков, поскольку в реальных условиях эксплуатации сегменты речи могут быть либо зашумлены, либо перемежаться с музыкальными фрагментами. В рамках второго блока осуществляется обучение акустических и лингвистических моделей речи. Данное обучение осуществляется на основе предварительно подготовленных баз данных. Третий блок позволяет сформировать гипотезы о наличии или отсутствии некоторых слов в некоторых временных рамках на основе полученных акустических и лингвистических моделей. В рамках четвертого блока осуществляется верификация найденных слов.

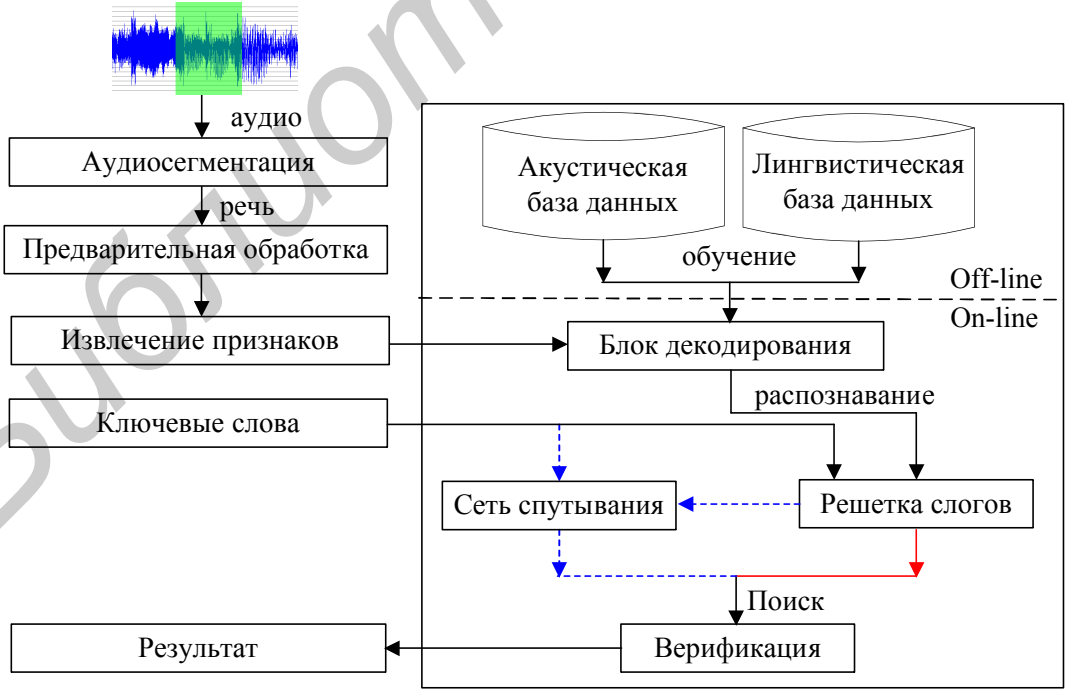

Рисунок 1 – Схема работы системы распознавания речи

2. Байесовский подход к задаче распознавания речи

Задачу РСР можно представить следующим образом [3, 17, 18, 19]. Пусть имеется ограниченный набор слов V, составляющий словарь распознавания. Представим речевой сигнал на входе системы в виде некоторой последовательности векторов признаков O. Система РСР принять решение о наличии в принятой реализации сигнала O слова $W, W \in V$.

Для решения этой задачи правомерно использовать Байесовский критерий минимума средних потерь. В этом случае критерий минимума средних потерь можно преобразовать к критерию максимума апостериорной вероятности (МАВ):

$$\overline{W} = \arg\max_V P(W \mid O) = \arg\max_V \frac{P(O,W)}{P(O)} = \arg\max_V P(O,W) = \arg\max_V P(W)P(O \mid W), \quad (1)$$

где $P(W)$ -вероятность слова W, $P(O)$ -вероятность последовательности векторов наблюдений O, $P(O \mid W)$ -условная вероятность того, что слово W реализуется в виде последовательности O.

Вероятность $P(O)$, очевидно, не играет роли при выборе гипотезы $\overline{W}$ в формуле (1), поскольку она постоянна для любого ключевого слова W.

Таким образом, при таком подходе для решения задачи распознавания необходимо:

- представить речевой сигнал в виде последовательности векторов признаков O;
- вычислить вероятность $P(W)$ для каждого слова из словаря $W \in V$;
- вычислить условную вероятность $P(O \mid W)$;
- используя $P(W)$ и $P(O \mid W)$ выбрать гипотезу в соответствии с решающим правилом (1).

3. Акустическое моделирование

Для акустического моделирования речи и вычисления $P(O \mid W)$ наиболее широко используется скрытая Марковская модель (СММ), которую можно определить следующим образом [3, 7, 8]:

- конечное пространство состояний $\{1,2,...,S\}$;
- пространство наблюдений $R^D = \{O\}$, где D-размерность вектора признаков;
- вероятности перехода между состояниями $a_{ij} = P(s_t = i \mid s_{t-1} = j)$;
- выходные распределения для состояний $b_j(.) = P(o_t \mid s_t = j)$.

Общая структура СММ представлена на рисунке 2.

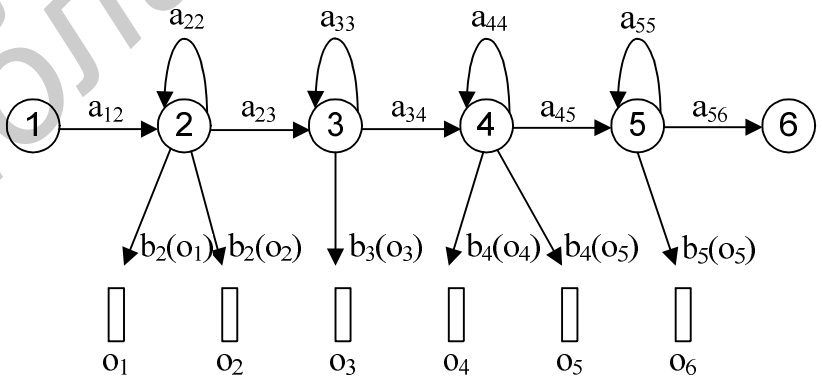

Рисунок 2 –Скрытая Марковская модель

Существует два основных подхода к акустическому моделированию на основе СММ [1, 10, 11]. Для систем РСР с небольшим словарем, как правило, для акустического моделирования создаются СММ для каждого слова из словаря. Для систем РСР с большим словарем такой подход практически неприемлем, и СММ для слов строится путем объединения СММ составных частей слов (фонем, слогов и т.д.). В этом случае гипотеза W представляет собой соединение СММ моделей акустических единиц, более мелких, чем слово. Такой подход позволяет моделировать одно и то же слово несколькими фонетическими последовательностями, что обеспечивает мощный инструмент для учета особенностей произношения слов разными людьми.

Для создания системы РСР важным этапом является выбор базовой единицы декодирования. Данный выбор должен быть осуществлен в соответствии с тремя основными принципами:

- гибкость: возможность составить из них другие фонетические и грамматические единицы;
- устойчивость: они должны сохранять устойчивость независимо от фонетического контекста;
- вычислительная сложность: компромисс между временем распознавания и сложностью модели.

Данные условия являются взаимно исключающими. Для обеспечения максимальной гибкости при построении акустической модели необходимо использовать фонетические единицы как можно меньшего размера. В то же время для максимальной устойчивости размер акустической единицы должен быть как можно больше, например, слово или словосочетание. Одновременно увеличение базовой акустической единицы приводит к значительному увеличению времени поиска ввиду увеличения количества вариантов. Задача выбора базовой акустической единицы состоит в выборе оптимального варианта с точки зрения одновременного удовлетворения условиям гибкости, устойчивости и вычислительной сложности.

Выбор базовой единицы декодирования находится в сильной зависимости от условий решаемой задачи и языка речи. Декодирование фонем в чистом виде допускает двойственное членение на слоги, что придает дополнительный элемент неустойчивости процессу акустического моделирования. В связи с этим более эффективным является использование дифонов и трифонов, представляющих собой контекстно-зависимые модели фонем. При акустическом моделировании на основе дифонов учитывается характеристики не только самой фонемы, но также и фонемы, находящейся слева или справа. Использование трифонов позволяет учесть одновременно как левый, так и правый фонемный контекст. Модель на основе дифонов и трифонов обладает высокой степенью устойчивости. Однако подход на основе трифонов имеет ряд существенных недостатков, связанных в первую очередь с их большим числом. При моделировании векторов наблюдений гауссовскими смесями число их компонентов варьируется от 8 до 16. В зависимости от количества выбранных трифонов общее количество оцениваемых параметров акустической модели языка составляет десятки и сотни миллионов. При условии ограниченности обучающих данных не всегда представляется возможным произвести качественную оценку параметров моделей трифонов. Кроме того, возможно появление ситуаций, когда нет соответствующего трифона для некоторого сочетания.

Проблема большого количества параметров и малого количества доступных данных для обучения является очень серьезной при разработке систем РСР, основанных на статистических методах распознавания. Для решения данной проблемы ранее использовался набор заранее определенных гауссовых распределений, и каждое состояние СММ моделировалось как набор весов этих распределений, который впоследствии сглаживался. Такого рода системы называются системами, основанными на связывании смесей. Высокая эффективность такого подхода была подтверждена экспериментально. Однако позднее была разработана методика сглаживания, основанная на связывании параметров, в частности, связывания состояний. Использование методик связывания в непрерывных СММ-системах приводит к значительному улучшению точности акустического моделирования.

4. Лингвистическое моделирование

Расчет апостериорной вероятности P(W), представляющей собой вероятность реализации некоторой последовательности слов W , можно произвести на основе использования априорных знаний о структуре, лингвистике и особенностях анализируемого языка. Существует целый ряд подходов к построению такого рода лингвистической модели, одним из наиболее широко распространенных является подход на основе n-граммной модели. Основная идея такого подхода состоит в том, что определение вероятности слова w_i производится с учетом вероятностей предшествующих n-1 слов. Определим апостериорную вероятность P(W) следующим образом*:*

$$P(W) = \sum_{i=1}^{N_W} P(w_i \mid w_{i-n+1},...,w_{i-1}), \quad (2)$$

где N_w - количество слов в последовательности *W*.

Основной проблемой лингвистического моделирования, как и при акустическом моделировании, является недостаточный объем обучающих данных. В связи с этим некоторые оценки вероятностей $P(w_i \mid w_{i-n+1},...,w_{i-1})$ могут оказаться равным нулю, что не отвечает реальному положению дел. В связи с этим необходимо проводить операции лифтинга и присваивать нулевым значениям вероятности минимальные ненулевые значения. Обычно используются биграммные или триграммные лингвистические модели, где n=2 или n=3 соответственно.

5. Декодирование

В процессе декодирования последовательности векторов наблюдений O происходит оценка апостериорной вероятности гипотез фонем. Для определения последовательности фонем наиболее широко используется алгоритм перемещающегося маркера, являющийся обобщением алгоритма Витерби для сети. Каждому узлу и связи в сети ставится в соответствие величина апостериорной вероятности, которая затем прибавляется к величине правдоподобия маркера, если они становятся частью анализируемого пути. Все возможные пути и меры их правдоподобия формируют сеть для распознавания.

Для построения системы РСР в качестве узлов такой сети можно использовать состояния СММ, тогда связь между узлами целиком определяется вероятностями перехода между состояниями СММ, а соответствующие значения правдоподобия равны логарифмическим вероятностям $L_i = \tilde{a}_{i,j}$ и $L_n = \tilde{b}_j(o_t)$ соответственно.

Пусть создана сеть декодирования, представляющая собой семантическую сеть, узлами которой являются все слоги, и возможен переход из конечного состояния одного слога в начальное состояние другого слога (рисунок 3). Во время декодирования каждому состоянию сопоставляется один маркер, в котором сохраняется частичное значение стоимости. Далее маркеры из каждого состояния передаются во все возможные последующие состояния, в каждом из которых сохраняется маркер с максимальной накопленной вероятностью. Для конечного состояния каждого слога СММ добавляется информация о вероятности лингвистической модели и штрафе, и сохраняется разметка слогов маркера с наибольшей вероятностью на текущий момент.

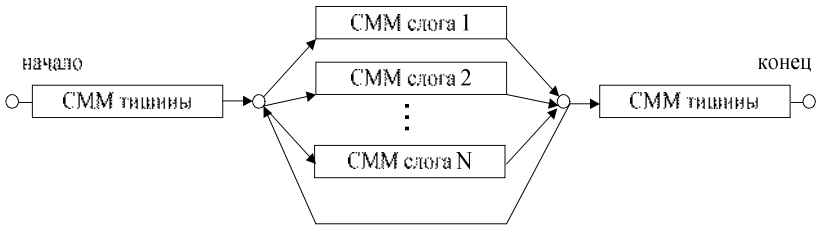

Рисунок 3 – Структура семантической сети для декодирования слогов

После того, как речевой фрейм полностью обработан, выполняется обратный ход алгоритма, позволяющий на основе маркеров восстановить последовательность слогов с наибольшим значением апостериорной вероятности. В результате этой процедуры формируется семантическая сеть возможных слогов, включающая в себя информацию о времени начала и конца каждого слога, значении акустической вероятности каждого пути и значение лингвистической модели. Результат распознавания цепочки слогов можно представить в виде семантической сети, включающей в себя различные варианты цепочки слогов. Для уменьшения количества вычислений можно использовать алгоритм сокращения количества путей, который исключает из дальнейшего рассмотрения пути с низкими значениями вероятности путем сравнения с некоторым заданным порогом. Данный алгоритм позволяет управлять процессом декодирования и получать результаты с разной точностью и вычислительной сложностью.

6. Компактное представление результатов декодирования

В результате выполнения процедуры декодирования слитной речи на основе акустической модели формируется набор путей через модель, каждому из которых соответствует свой набор маркеров и своя вероятность. Существует две возможности представления результатов декодирования в компактном виде: в виде N-best гипотез, и в виде семантической сети слогов [13, 14].

На основе алгоритма перемещающегося маркера с каждым состоянием СММ j в момент времени t сопоставляется некоторый маркер - структура, содержащая значения и состояния для наиболее вероятного пути, оканчивающегося в состоянии j для всех наблюдений до момента времени t. Если некоторому состоянию соответствует много маркеров, то можно сохранять как все маркеры, так и только маркер с наибольшим значением вероятности для уменьшения вычислительной сложности. После обработки всех речевых данных по информации, хранящейся в маркерах, можно восстановить наиболее вероятные последовательности слов. Таким образом, информация, хранящаяся в маркерах, дает компактное представление множества наиболее вероятных гипотез, называемое сетью. На основании семантической сети можно сгенерировать список, состоящий из N гипотез с наибольшими вероятностями, называемый *N-best* списком.

Однако использование *N-best* списка является эффективным только в том случае, если число N имеет порядок десятков и сотен, поскольку если используемые для декодирования модели недостаточно оптимальны, то правильная последовательность с большой долей вероятности не попадает в число *N-best* гипотез при небольшом числе N. Кроме этого, *N-best* список, как правило, обладает большой избыточностью, многие последовательности могут отличаться только вариациями одного из слов. Общее число гипотез растет экспоненциально с увеличением длины высказывания, и в этих условиях выбор среди *N-best* гипотез становится все менее эффективным.

В связи с этим были разработаны семантические сети и графы слов в качестве компактного представления альтернативных гипотез для замены *N-best* списков. Структура и алгоритмы построения сети достаточно сложны, однако они позволяют очень эффективно и компактно представить множество гипотез. Простейшая такая сеть представляет собой *N-best* список. Графы слов, с другой стороны, представляют собой конечный автомат состояний, в котором каждое слово описывается дугой слов.

Основное преимущество использования семантической сети перед списком *N-best* заключается в том, что для нее не происходит дублирования расчета вероятностей для общих подпоследовательностей слов, как это характерно для *N-best* метода, в связи с этим сеть намного эффективнее с точки зрения вычислительной сложности.

Исследования показали, что компактное представления результатов декодирования в виде списка *N-best* или сети не накладывает ограничений на сложность применяемых лингвистических и акустических моделей. Сеть слов может быть легко расширена на случай использования лингвистических моделей. Эффективность генерации семантической сети была значительно повышена за счет использования лексических деревьев. Позже была произведена интеграция лингвистической модели непосредственно в семантическую сеть, и получена более точная и эффективная структура.

Семантическая сеть слов может быть получена на основе алгоритма динамического программирования с использованием биграммных и триграмных моделей. Однако для генерации семантической сети наиболее удачными признаны СММ. Генерация и использование сети на основе СММ позволяет значительно улучшить эффективность и точность системы распознавания речи по сравнению с применением только СММ. Структура сети позволяет генерировать и использовать ее для различных базовых акустических единиц-слов, слогов, фонем и т.д.

7. Структура семантической сети слогов

Для компактного представления результата фонетического декодера вводится понятие сети следующим образом. Семантическая сеть $L = (N, V, n_{start}, n_{end})$ представляет собой направленный непериодический граф, где N – множество узлов, V – множество дуг, каждая из которых представляет гипотезу единицы декодирования (слово, слог, фонема), $n_{start}, n_{end} \in N$ начальный и конечный узлы. Каждому узлу сети $n \in N$ соответствует свое время $t(n)$. Представим дуги в виде четырехугольников $(s[v], e[v], w[v], p[v])$, где $s[v], e[v] \in N$ – соответственно начальный и конечный узлы дуги v, $w[v]$ – базовая единица декодирования, $p(v)$ – акустическая вероятность, соответствующая речевому фрагменту $O_{s(v)}^{e(v)}$, начинающемуся в $s[v]$ и заканчивающемуся в $e[v]$, $p(v) = P(O_{s(v)}^{e(v)} \mid w(v))$. В этом случае предложение можно представить как часть графа $[w; s, e]_1^M = [w_1; s_1, e_1], \ldots, [w_M; s_M, e_M]$, где 1-начальный момент времени, М- последний момент времени.

На рисунке 4 представлена структура семантической сети слогов.

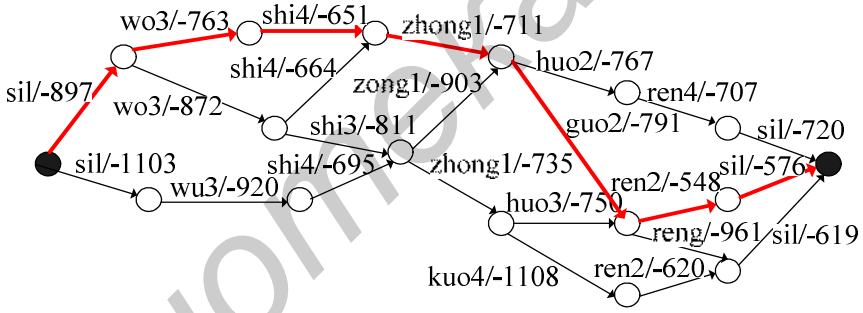

Рисунок 4 – Структура семантической сети слогов

На рисунке 4 sil обозначает тишину, правильный путь декодирования обозначен жирной линией. По сути, сеть представляет сжатое пространство декодирования, включающее в себя значимую информацию декодирования в процессе распознавания, узлы пересечения определяют конкурентные версии, а полный путь из начального узла до конечного может восприниматься как последовательность фонетических единиц гипотетического предложения.

8. Вычисление апостериорной вероятности слов

В качестве единичной характеристики для определения правдоподобия той или иной гипотезы распознавания речевой последовательности слов может выступать ее апостериорная вероятность, которая легко может быть вычислена на основе содержимого семантической сети. Исходя из этого утверждения, введем меру достоверности следующим образом.

Представим строку слогов, обозначающих некоторое слово W, как $v_1 v_2 \ldots v_K$, тогда апостериорная вероятность $P(W \mid O)$ будет иметь вид:

$$P(W \mid O) = P(v_1 v_2 \ldots v_K \mid O), \qquad (3)$$

где O- последовательность векторов признаков речевого сигнала. Пусть V^* - множество всех гипотетических путей, включающих строку слогов $v_1v_2...v_K$, тогда апостериорную вероятность слова W можно представить как

$$P(W \mid O) = \sum_{v_1v_2...v_K \subset V^*} P(v_1v_2...v_K) \mid O), \quad (4)$$

где V^* - множество всех гипотетических путей, включающих последовательность $w = v_1v_2...v_K$.

В этом случае вероятность некоторого слова совпадает с накопленной апостериорной вероятностью всех возможных путей. Обозначим путь сопоставления $w = v_1v_2...v_K$, тогда апостериорную вероятность (3) можно записать как $P(w \mid O)$. Необходимо вычислить апостериорную вероятность каждого сопоставленного пути, и после накопления можно получить апостериорную вероятность слов. Апостериорную вероятность можно представить следующим образом:

$$P(w \mid O) = \frac{P(O, w)}{P(O)}. \quad (5)$$

При вычислении (3) необходимо рассматривать все возможные пути, количество которых может быть очень большим. Для быстрого вычисления $P(w \mid O)$ наиболее удачным является алгоритм прямого-обратного хода [15, 16].

Для уменьшения вычислительной сложности для сети слогов используется специальная процедура сокращения избыточности сети, позволяющая удалить пути с низкой вероятностью

Очевидно, что данный алгоритм устранения избыточности не является лучшим в общем смысле, поскольку существует возможность устранения пути, обеспечивающего глобальный максимум вероятности. Вероятность появления такой ошибки напрямую зависит от величины порога - чем выше порог, тем выше вероятность ошибки, но ниже вычислительная сложность. Вопрос о том, как получить эффективный компромисс между вероятностью ошибочного распознавания и временем распознавания, является предметом экспериментального исследования. Кроме этого, алгоритм устранения избыточности может быть реализован путем ограничения количества путей на каждом уровне анализа. В этом случае в качестве порога избыточности выступает некоторое фиксированное число.

Для вычисления апостериорных вероятностей и обеспечения эффективности их использования в качестве критерия достоверности необходимо выполнить процедуру масштабирования. Если вероятности акустической модели не масштабированы надлежащим образом, то при суммировании во всех вышеприведенных выражениях будут преобладать только несколько гипотез слоговой сети ввиду очень большого динамического диапазона акустических значений (а соответственно, и отрицательных значений логарифма ненормализованных акустических вероятностей). Акустические признаки обладают высокой степенью изменчивости, и оценка их дисперсии в общем случае является сложной задачей. Ввиду этого предлагается использовать процедуру масштабирования.

Оба параметра масштабирования α и β для акустической и лингвистической моделей соответственно необходимо оценить с использованием обучающего набора. Во время выполнения алгоритма прямого-обратного хода, все вероятности лингвистической модели масштабируются на коэффициент β, а все вероятности акустической модели на коэффициент α. В результате получим:

$$p(w_1^M \mid o_1^T) = p(o_1^T \mid w_1^M)^{\alpha} \cdot p(w_1^M)^{\beta} \quad (6)$$

Все остальные выражения модифицируются соответствующим образом.

Эффективность использования апостериорной вероятности в качестве критерия достоверности, сильно зависит от правильного выбора коэффициента акустического масштабирования α . Оптимальная величина для коэффициента масштабирования модели языка β , когда оптимизация α и β производится на обучающем наборе, не постоянная, но близкая к 1.0. Это говорит о том, что единица будет ожидаемым значением, если вероятности лингвистической модели нормализованы.

9. Семантическая сеть спутывания

Семантическая сеть спутывания является компактной формой записи конкурирующих временных гипотез, полученных в результате декодирования речи. Она широко применяется в различных задачах обработки речи. Сеть спутывания создается путем выравнивания всех путей сети фрагментов речи, т. е. приведения речевой семантической сети к линейному виду. На данный момент не разработано эффективных методов выравнивания речевых сетей ввиду наличия ряда проблем. Первая из них связана со сложностью структуры речевой сети, которая затрудняет расчет шкалы выравнивания. Второй проблемой является то, что низкие вероятности появления взаимоисключающих гипотез в сети могут повлиять на эффективность процесса декодирования в полученной сети спутывания. В качестве решения первой проблемы возможно использовать методы сегментации сети, а для решения второй проблемы используется переоценка лингвистической вероятности появления некоторого речевого высказывания с помощью триггерных моделей.

Семантическая сеть спутывания, получаемая путем извлечения информации из словесных структур, дает более ясное представление обо всех конкурирующих гипотезах, и ее использование поможет улучшить точность декодирования речи. При использовании стандартного метода оценки максимума апостериорной вероятности (МАВ) система декодирования оценивает последовательность слов, представляющую собой путь с максимальной апостериорной вероятностью, полученный при заданных акустической и лингвистической моделях. Однако данный метод не всегда приводит к минимизации коэффициента ошибок слов(КОС)- отношение суммы ошибочно замененных, вставленных и удаленных слов к общему числу слов в выражении.

Вместе с тем сеть спутывания позволяет явно минимизировать КОС путем извлечения гипотез с максимальными апостериорными вероятностями. Таким образом, происходит приведение многоуровневых словесных структур к компактному виду, а так же замена глобального поиска по огромному набору гипотез-предложений на локальный поиск по небольшому множеству вероятных словесных гипотез.

При декодировании при помощи метода МАВ поиск ключевых слов осуществляется путем выбора последовательности слов $\Omega^{\max}$ с максимальной апостериорной вероятностью для заданного акустического наблюдения O:

$$\Omega^{\max} = \arg\max_{\forall \Omega_i \in \Omega} P(\Omega_i \mid O), \tag{7}$$

где O – акустическое наблюдение, Ω_i – возможная последовательность слов (предложение); Ω – множество всех возможных предложений.

Вычислительную сложность процедуры можно значительно снизить, если структуру семантической сети слов привести к структуре сети спутывания. Кроме этого, в отличие от метода МАВ такой подход дает возможность для анализа рядов конкурирующих гипотетических слов в предложении. Еще одно отличительное достоинство сети спутывания – это ее линейная структура, что позволяет легко вводить в сеть дополнительные ограничения. В настоящее время сеть спутывания является самой компактной формой записи гипотетического результата декодирования слитной речи.

Семантическая сеть спутывания является компактным видом представления множества гипотез последовательности слогов, полученных из структур сети. На основе теории метода максимальных апостериорных вероятностей следует, что последовательность слогов, получившая максимальную вероятность в процессе декодирования, сводит к минимуму КОС.

Однако на практике между критерием эффективности декодирования и критерием минимума КОС существует противоречие, минимум КОС не всегда обеспечивается последовательностью слогов с минимальной вероятностью.

Сеть спутывания представляет собой особый вид словесной структуры (рисунок 5), которую можно рассматривать как направленный непериодический граф.

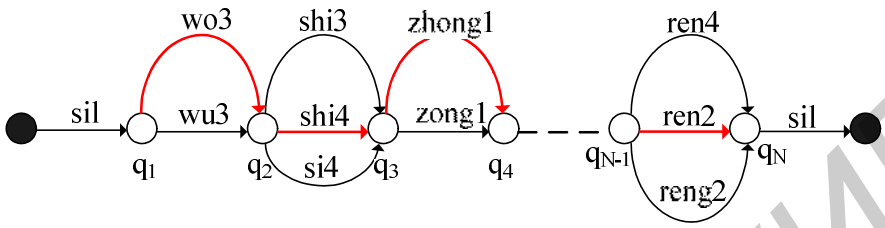

Рисунок 5 – Структура сети спутывания

Обозначим семантическую сеть спутывания как $F=(Q,q_1,q_N,E)$, где $Q=\{q_1,...,q_N\}$ - множество всех узлов, q_1 и q_N - начальный и конечный узлы сети, N – количество узлов, E – множество всех дуг. Каждая дуга определяется четырьмя параметрами $e=(q_e^s,q_e^f,w_e,g_e)$, где $q_e^s,q_e^f\in Q$ - начальный и конечный узлы дуги e (если $q_e^s=q_i$, то $q_e^f=q_{i+1}$), w_e - индекс дуги, а $g_e\in(0,1)$ задает апостериорную вероятность возникновения дуги e. Сеть спутывания, состоящая из N узлов, включает N–1 множеств спутывания $Z_i, i=1,...N-1$, каждое из которых определяется множеством дуг, для которых совпадают начальный и конечный узлы. Если в множестве спутывания для некоторого узла существуют дуги с одинаковыми индексами, то они объединяются в одну дугу.

Основной задачей генерации сети спутывания является обеспечение минимального КОС, что является основной задачей при создании систем ПКС.

Использование сети спутывания имеет целый ряд преимуществ по сравнению с обычной сетью. Прежде всего, декодирование при помощи семантической сети дает решение только на основе полной гипотезы, а в случае сети спутывания может быть проверен ряд конкурирующих гипотетических слов. Во-вторых, сеть спутывания имеет линейную структуру, которую более удобно дополнять информацией от различных других модулей. Уникальные характеристики сети спутывания обеспечили ей широкое применения в различных областях, таких как распознавание речи и языка, понимание взаимосвязи между словами, обнаружение границ предложений, оценка мер доверительности.

После проведения процедуры выравнивания и построения сети спутывания, можно найти путь предложения с минимальным КОС. Как следует из приведенных выше рассуждений, при использовании такой процедуры генерации сети и путь с наименьшим КОС получается при выборе для каждого узла дуги с наибольшим значением апостериорной вероятности.

10. Улучшение лингвистической оценки речевых выражений на основе триггеров

В любой произнесенной речи всегда можно выделить много контекстных взаимосвязей между словами [4]. Человек быстрее и лучше воспринимает тесно связанную пару слов, чем слабо связанную. Контекстная взаимосвязь может наблюдаться между двумя существительными (например «день / ночь»), между прилагательным и существительным (например «громкий / голос»), в устойчивых выражениях (например «обращать / внимание») и др. Такие взаимосвязи можно получить из набора предварительно заданных данных и затем использовать для устранения неоднозначностей в распознаваемом предложении.

На данный момент преобладают простые n-граммные модели языка, которые могут учитывать близкую контекстную зависимость в пределах окна размером n слов (обычно n=3). Однако большинство зависимостей может оказаться за пределами этого окна. Таким образом, в *n*-граммных моделях фиксируются зависимости только на малых дистанциях.

Основываясь на вышесказанном, в качестве основной концепции для извлечения ассоциативной информации из связанной пары слов на малых и больших дистанциях для поиска ключевых слов было предложено использовать модель триггеров (рисунок 6). Структуру $(A_0 \to B)$ можно рассматривать как триггер, если инициируемое слово A_0 тесно связано с инициируемым словом B. Когда слово A_0 встречается в документе, оно инициирует слово B, вызывая переоценку вероятности его появления.

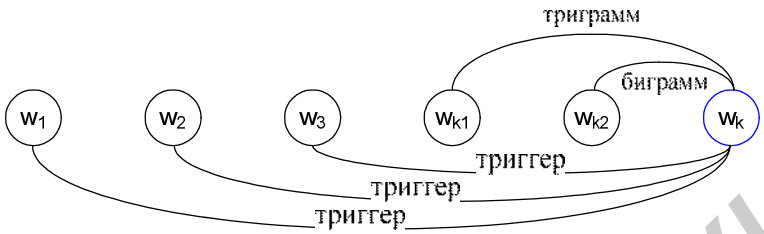

Рисунок 6 – Триггерная модель для некоторого слова

w_k , $w_1, w_2, w_3, w_{k1}, w_{k2}$ - предшествующие слова

Даже если ограничиваться рассмотрением парных триггеров, то количество таких взаимосвязей может быть очень большим. Поэтому для моделирования языка на основе триггеров уместно выбирать наиболее значимые триггеры.

11. Недостатки и пути совершенствования современных систем распознавания слитной речи

Несмотря на все прилагаемые усилия, точность современных систем распознавания слитной речи остается на недостаточно высоком уровне. Это обусловлено несколькими основными причинами.

Во-первых, речевой сигнал обладает значительной акустической изменчивостью, выражающейся в том, что даже один и тот же человек в разные моменты времени одно и то же слово говорит по разному. Акустические характеристики речи еще больше различаются для различных людей с разными региональными акцентами, особенностями произношения, дефектами речи. Поскольку все современные системы распознавания речи основаны статистических моделях, то для качественной оценки параметров при таком разбросе акустических характеристики требуется огромное количество обучающих данных, недостижимое в реальных условиях. Ситуация ухудшается еще за счет того, что добавление любого нового диктора в обучающую выборку не обязательно ведет к улучшению точности оценки параметров, а наоборот, способно ее ухудшить.

Во-вторых, использование стандартных методов лингвистической обработки последовательностей акустических символов, полученных на выходе декодера, не позволяет в полной мере охватить все возможные варианты речевых сообщений ввиду их огромного количества. В связи с этим приходится применять различные методы сокращения количества возможных последовательностей, что приводит к увеличению количества ошибок. Кроме того, стандартные методы не позволяют в полной мере учесть все лингвистические и семантические связи между словами, что в значительной степени облегчило бы распознавание речевых высказываний.

В-третьих, имеющиеся в настоящее время средства РСР не позволяют выполнить основную задачу речевых интерфейсов - понять смысл речевого высказывания. Это наиболее существенный недостаток существующих и широко применяемых подходов. Без решения задачи понимания речевых высказываний невозможно построение интеллектуальных систем нового поколения.

Успешное решение каждой из перечисленных выше проблем возможно только при использовании методов искусственного интеллекта, в частности, методов семантического анализа, описанных выше.

Библиографический список

1. Бовбель, Е.И. Скрытые марковские модели и машины на опорных векторах: от теории к практике : пособие для студентов / Е.И. Бовбель, И.Э. Хейдоров, Ю.В Пачковский ; Белорус. гос. ун-т. – Минск, 2008. – 130 с.
2. Лобанов, Б.М. Речевой интерфейс интеллектуальных систем : учеб. пособие / Б.М. Лобанов, О.Е. Елисеева ; Белорус. гос. ун-т информатики и радиоэлектроники. – Минск, 2006. – 151 с.
3. Рабинер, Л.Р. Скрытые марковские модели и их применение в избранных приложениях при распознавании речи : обзор / Л.Р. Рабинер // Тр. ин-та инженеров по электронике и радиотехнике. – 1989. – Т. 77, № 2. – С. 86–120.
4. Хейдоров И.Э., Сорока А.М., Янь Цзиньбинь. Поиск ключевых слов в слитной речи для современных систем обработки аудиосигналов // мон., РИВШ Минск, –2010.
5. A new verification-based fast-match for large vocabulary continuous speech recognition / M. Afify [et al.] // IEEE Trans. on Audio, Speech a. Language Processing. – 2005. – Vol. 13, № 4. – P. 546–553.
6. Automatic recognition of keywords in unconstrained speech using hidden Markov models / J.G. Wilpon [et al.] // IEEE Trans. on Audio, Speech a. Language Processing. – 1990. – Vol. 38, № 11. – P. 1870–1878.
7. Beyond ASR 1-best: using word confusion network in spoken language understanding / D. Hakkani-Tur [et al.] // Computer Speech a. Language. – 2006. – Vo1. 20, № 4. – P. 495–514.
8. Bourlard, H. Recognition and rejection performance in word spotting systems / H. Bourlard, D.H. Bart, O. Boite // IEEE International conference on acoustics, speech and signal processing, Adelaide, 19–22 Apr. 1994. – Adelaide, 1994. – Vol. 1. – P. 373–376.
9. Bovbel, E.I. The joint speech/video signal processing based on the autoregressive hidden Markov models and neural networks / E.I. Bovbel, P.D. Kukharchik, I.E. Kheidorov // Intelligent signal processing : proc. of IEEE Intern. workshop on intelligent signal processing , Budapest, 4–7 Sept. 1999. – Budapest, 1999. – P. 45–47.
10. Bridle, J.S. An alphanet approach to optimizing input transformations for continuous speech recognition / J.S. Bridle, L. Dodd // IEEE International conference on acoustics, speech and signal processing, ICASSP 1999, Phoenix, 14–18 March 1999. – Phoenix, 1999. – Vol. 1. – P. 227–280.
11. Juang, B.H. Mixture autoregressive hidden Markov models for speech signals / B.H. Juang, L.R. Rabiner // IEEE Trans. on Audio, Speech a. Language Processing. – 1985. – Vol. 33, № 6. – P. 1404–1413.
12. Kamppari, S.O. Word and phone level acoustic confidence scoring / S.O. Kamppari, T.J. Hazen // IEEE International conference on acoustics, speech and signal processing, ICASSP 2000, Istanbul, 5–9 June 2000. – Istanbul, 2000. – Vol. 3. – P. 1799–1802.
13. Kenny, Ng. Subword-based approaches for spoken document retrieval / Ng. Kenny, V.W. Zue // Speech Communication. – 2000. – Vol. 32, № 3. – P. 157–186.
14. Keyword spotting based on syllable confusion network / P.Y. Zhang [et al.] // 3rd International conference on natural computation ICNC 2007, Hainan, 24–27 Aug. 2007. – Hainan, 2007. – P. 656–659.
15. Mangu, L. Finding consensus in speech recognition: word error minimization and other applications of confusion networks / L. Mangu, E. Brill, A. Stolcke // Computer Speech a. Language. – 2000. – Vol. 14, № 4. – P. 373–400.
16. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods / J. Platt // Advances in large margin classifiers / ed.: A.J. Smola [et al.]. – Cambridge, 1999. – P. 61–74.
17. Rohlicek, J.R. Continuous hidden Markov modeling for speaker-independent word spotting / J.R. Rohlicek, W. Russell // IEEE International conference on acoustics, speech and signal processing, ICASSP 2005, Glasgow, 22–25 May 1989. – Glasgow, 1989. – Vol. 1. – P. 627–631.
18. Rose, R.C. Keyword detection in conversational speech utterance using hidden Markov model based continuous speech recognition / R.C. Rose // Computer, Speech a. Language. – 1995. – № 9. – P. 309–333.
19. Rosenfeld, R. A Maximum entropy approach to adaptive statistical language modeling / R. Rosenfeld // Computer Speech a. Language. – 1996. – Vol. 10, № 2. – P. 187–228.
20. Wright, J.H. A consolidated language model for speech recognition / J.H. Wright, G.J.F. Jones, H. Lloyd-Thomas // Proceedings of EUROSPEECH–93, Berlin, 21–23 Sept. 1993. – Berlin, 1993. – Vol. 2. – P. 977–980.