

ОБЗОР МЕТОДОВ ПОНИМАНИЯ РЕЧИ И ТЕКСТА

Ю. А. Косарев, И. В. Ли, А. Л. Ронжин, Е. А. Скиданов¹, J. Savage²

¹Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия В.О., д.39
kosarev@mail.iias.spb.su

²University of Mexico, UNAM. Mexico D.F. C.P. 04510

УДК 381.3

Ю. А. Косарев, И. В. Ли, А. Л. Ронжин, Е. А. Скиданов, J. Savage. **Обзор методов понимания речи и текста** // Труды СПИИРАН. Вып. 1, т. 2. — СПб: СПИИРАН, 2002.

Аннотация. *Предлагается краткий обзор существующих методов понимания речи и текста на основе анализа публикаций в научных журналах и трудах ведущих конференций по речевой тематике. Выделяются два различных подхода к проблеме понимания языка: 1) на основе исчисления высказываний и 2) на основе распознавания смысла (речевого намерения). Приводятся доводы в пользу интегральной парадигмы обработки речи, разрабатываемой СПИИРАН по сравнению с широко известными подходами. — Библ. 97 назв.*

UDC 381.3

Yu. A. Kosarev, I. V. Lee, A. L. Ronzhin, E. A. Skidanov, J. Savage. **Review of speech and text understanding methods** // SPIIRAS Proceedings. Issue 1, v. 2. — SPb: SPIIRAS, 2002.

Abstract. *Brief review of existent methods for speech and text understanding is presented based on analysis of publications in science journals and leading conferences on speech research. Two different approaches to language understanding problem are pointed: 1) approach based on propositional calculus and 2) approach based on meaning recognition (speech intention). The arguments for speech processing paradigm elaborated in SPIIRAS are presented in comparison with conventional approaches — Bibl. 97 items.*

1. Введение

Эта статья кратко описывает историю развития исследований в области понимания языка и речи. Рассматриваются как традиционные лингвистические подходы к данной проблеме, так и последние направления, возникшие на стыке лингвистики, искусственного интеллекта и информатики. Приводится анализ состояния проблемы распознавания речи и высокоуровневой обработки. Делается попытка оценить меру перспективности существующих подходов к проблеме.

В главе 2 приводится краткий обзор известных подходов к высокоуровневой обработке языка (текста). В главе 3 проводится анализ состояния исследований автоматического понимания речи. В главе 4 описывается подход СПИИРАН к проблеме речевого диалога [1].

Здесь необходимо сразу оговориться, что понимание языка и речи здесь рассматривается в рамках проблемы человеко-машинного диалога. Диалоговые системы первых поколений обычно базировались на лингвистических автоматах, назначение которых было проверять истинность/ложность входных высказываний. Однако сейчас все большую популярность приобретают модели коммуникации, позволяющие рассмотреть совместно все основные звенья информационного процесса, такие как: оператор, вычислитель, предметная область, а также рассмотреть взаимодействие языковых и неязыковых знаний. На рис. 1 показана модель человеко-машинной коммуникации,



Рис. 1. Модель человеко-машинной коммуникации

где показаны основные компоненты: человек, обладающий необходимой априорной информацией о языке данной предметной области; он также учитывает текущую информацию о изменениях в предметной области; компьютер, выполняющий распознавание/понимание речи и исполняющий задачи диалога (выдачу справочной информации, перевод речи на другой язык, воздействие на какие-либо механизмы и др.); канал связи (простая проводная линия, телефонный канал, радиоканал и др.); предметная область, это сфера, на которую направлены результаты диалога (создаваемый файл-перевод с исходного языка, управляемый робот, телефонная автоматическая справочная служба и др.), и она же — источник нужной текущей информации для человека.

Для того чтобы исключить влияние ошибок распознавания/понимания, в системе должна быть предусмотрена проверка (визуальная или слуховая) правильности распознавания, после чего оператор либо подтверждает правильность распознавания смысла либо выбирает другую альтернативу из предложенного списка.

2. Краткий обзор лингвистических подходов к пониманию языка (текста)

В данном разделе мы предлагаем краткий обзор существующих методов понимания речи и текстов, а также обосновываем доводы в пользу парадигмы обработки речи СПИИРАН по сравнению с широко известными подходами.

В области высокоуровневой обработки речи и текста существуют две принципиально разные парадигмы: исчисление смысла высказывания из распознанных слов и определение речевого намерения говорящего. Первая уходит корнями в языковые лингвистические теории, такие как теория семантических падежей Филлмора, теория концептуальных зависимостей Шенка, теория семантических сетей и др. Все они направлены на исчисление смысла высказывания из входных слов по определенным правилам. Эта парадигма основана на предположении, что язык является логическим объектом, а мышление — логическим исчислением. Модели, основанные на данном подходе, обладают рядом слабых сторон. Во-первых, сторонники данного подхода склонны углубляться в

сам процесс исчисления высказывания в ущерб коммуникативной цели речевого диалога. Во-вторых, как правило, исчисление высказывания основано на последовательном разборе, и любая ошибка влечет за собой неверное решение, т.е. данный подход не учитывает вариативности языка, его неоднозначности и зависимости от контекста и внешних условий. Данная парадигма рассматривает язык как бы в одной плоскости, в основном грамматической. В действительности его надо видеть во всех необходимых для понимания аспектах, таких как семантический, ассоциативный, статистический, ситуативный, контекстуальный и т.д. Взвешенный учет всех этих аспектов позволяет создавать действительно эффективные системы.

Вторая парадигма возникла в противовес первой, она предполагает, что главной целью речевой коммуникации является передача речевого намерения, а главной целью слушающего является распознавание этого намерения, т.е. смысла (а не оценка типа истинно/ложно). Эта парадигма хорошо согласуется с теорией речевых актов [2], она приобретает все большую популярность. Кроме того, данная парадигма позволяет использовать аналогии из теории коммуникаций, т.е. использовать широко известный принцип дешифрации избыточных сообщений «по минимуму отклонений».

Вначале мы рассмотрим вышеуказанные лингвистические теории, их достоинства и недостатки, обсудим наиболее распространенный способ представления знаний в этих теориях, т.е. фреймовые структуры. Затем мы представим ряд альтернативных подходов к высокоуровневой обработке языка и речи, которые хорошо увязываются с интегральной парадигмой понимания (подход СПИИРАН). Опираясь на анализ публикаций трудов ведущих международных конференций в области обработки языка и речи за последние годы, мы попытаемся оценить, насколько оправдано использование известных подходов к обработке текста для речевых диалоговых систем.

2.1. Язык и искусственный интеллект

Здесь важно отметить связь исследований языка и речи с проблемой искусственного интеллекта. Известно, что естественный язык по своей сложности эквивалентен интеллекту человека. Таким образом, все сложности познания человеком самого себя и внешнего мира присутствуют в полной мере и при исследовании речемыслительной деятельности человека. Именно поэтому попытки смоделировать отдельные аспекты этой деятельности немедленно вводят нас в проблематику искусственного интеллекта (ИИ), который, как известно, имеет главной целью заимствование важнейших свойств живых систем для создания перспективных систем автоматической обработки и передачи информации

Существует, правда, и другая точка зрения, согласно которой, при создании работающих систем ИИ, не нужно опираться на речемыслительные прототипы человека, а следует искать оригинальные математические и инженерные решения, наращивая при этом, с одной стороны, логические и вычислительные ресурсы систем, а с другой — расширяя базы данных и знаний. Сторонники этой точки зрения ссылаются на примеры изобретения колеса или винтового самолета. Аналогии эти не слишком убедительны. Природными прототипами колеса, на которые мог опереться первобытный изобретатель, являются катящиеся по откосу камень или бревно, а планер самолета имитирует парящий полет птицы. Винт самолета, действительно иллюстрирующий редкий случай

отсутствия природного прототипа, представляет собой, тем не менее, устройство, которое нельзя сравнивать по сложности с речемыслительной системой человека, которую природа создавала многие десятки, а точнее сотни тысяч лет. В эти процессы вовлечены миллиарды нейронов, объединенных сложнейшими связями. Поэтому попытки создания систем искусственного интеллекта без учета базовых свойств живых прототипов, представляются наивными и самонадеянными, во всяком случае, они пока не принесли сколько-нибудь ощутимых результатов.

Известно, что именно лингвистическая проблематика послужила мощным стимулом для таких ключевых разделов теории искусственного интеллекта, как марковские модели, проблема неопределенности исходных данных, породившая теорию нечетких множеств и лингвистической переменной, функции доверия и т.п. Говоря о связи речевой проблематики с теорией ИИ надо подчеркнуть, что здесь также решается «вечная» проблема искусственного интеллекта (ранее она решалась также и в рамках кибернетики) *о построении надежных систем из ненадежных компонентов.*

2.2. Две парадигмы языка

Анализируя исследования в области обработки текстов и последующие исследования в области речи можно обнаружить две принципиально различные парадигмы языка.

Жестко логическая парадигма. Странники этой парадигмы считают, что язык человека представляет собой строго системное построение, генетически унаследованное каждым ребенком от его родителей, неспособное к постепенному эволюционированию. Эти мысли обнаруживаются еще у древних греков (Сократ, Платон), далее в Средние века (Лейбниц, а также есть масса последователей и в наши дни). Эти идеи были развиты американским логиком и лингвистом Н.Хомским [3]. Согласно их концепции в основе каждого естественного языка лежит жесткая универсальная система, использующая исчисление высказываний и охватывающая единый набор врожденных идей, передающихся генетически. Эти идеи составляют основу так называемой «структурной лингвистики» (СЛ).

В начале 60-х гг. многие математики, программисты и лингвисты увлеклись перспективой построения искусственного разума на основе описанной выше концепции. К сожалению, это направление не дало ожидаемых результатов, не позволило создать ни одной работающей в промышленном режиме системы ИИ, которая бы опиралась на вышеописанные идеи.

В противовес привычному принципу СЛ, допускающему только две категории текстов (истинно/ложно), стало очевидным, что письменные и звучащие фразы, которые вообще существуют в обиходе, могут быть в принципе разделены на три множества. Первое — это идеальные, предложения, соответствующие всем правилам и нормам языка. Второе — неточные, но потенциально понятные. Здесь возникла целая индустрия нахождения ошибок, алгоритмы, которые могут исправлять ошибки на основе статистических, или синтаксических закономерностей. И есть просто непонятные фразы. Эта точка зрения хорошо коррелируется с результатами теории информации, теорией передачи избыточных сообщений, где аналогичным образом существуют безошибочные сообщения, ошибки, которые потенциально исправимы, и неисправимые ошибки.

Эта классификация текстов в жизни обычно опирается на модель предметной области, в которой происходит процесс коммуникации.

Приводились данные [4], что более 25% текстов, которые пользователь вводит в компьютер, содержат различные ошибки, которые компьютер затем «выплевывает», и редактирование повторяется. Но в живой речи, мы ведь не поправляем, друг друга без конца, мы понимаем смысл и движемся в диалоге дальше, то есть постоянно имеем дело с разного рода некорректностями, которые наш интеллект обходит удивительно просто, чаще всего не замечая их.

Применение жестко логической парадигмы языка в системах ИИ наталкивается на целый ряд противоречий. Например, «истинность и ложность» в принципе не коррелируются с «понятностью и непонятностью». Так если речь не совсем корректна и «истинна», но понятна, то она в этом смысле правильна, так как позволяет реализовать основную, коммуникативную функцию языка.

Другое противоречие состоит в том, что из подмножества равноценных предложений, конструируемых структурно-лингвистическими алгоритмами, не существует принципиальной возможности выбрать некоторый оптимальный вариант. А это противоречит психологии речевой деятельности: например, известно, что переводчик, постоянно выбирает лучший с его точки зрения вариант.

Далее такое противоречие. Чем больше видов априорной информации используется в процессе понимания по строгим правилам, тем хуже статистическая точность и устойчивость процесса. Но это не согласуется с данными о психологии речевой деятельности.

Обычно при СЛ-подходе совершенно игнорируется прагматическая информация, прежде всего, знание ситуаций предметной области. И лингвисты, и психологи знают, что это обязательный компонент понимания у человека. Но структурная лингвистика обходится без модели внешнего мира и, в итоге не может дать необходимые знания для создания приемлемой модели речевого общения.

Жестко-структурные модели не могут объяснить удивительную устойчивость речевой коммуникации людей по отношению к неблагоприятным факторам и, тем более, служить базой для создания перспективных систем искусственного интеллекта, включая разнообразные системы устного диалога.

Таким образом, допущение, по которому речемыслительная деятельность человека, как и все его поведение, могут быть полностью формализованы в терминах единой системы правил, оказывается неверным с точки зрения современного этапа развития лингвистики, информатики и теории ИИ [5]. Этим и объясняется постепенный отказ многих разработчиков систем ИИ от жестко-структурной парадигмы и поиск новых альтернативных математических, лингвистических и инженерных решений, таких как стохастические модели языка, стохастические грамматики, ассоциативные модели языка и др. Поддержка этого направления высказывается со стороны философов [6].

Отмеченные особенности языка представляют как бы новую философию речевого обмена, которая согласуется с одной стороны, с методами количественной лингвистики (КЛ), а с другой стороны, с основными положениями теории информации. Согласно этой аналогии суть речевого обмена состоит в том, что говорящий превращает свое речевое намерение (смысл) в устную форму (то есть, кодирует исходную информацию), а слушатель извлекает смысл (то есть, декодирует сообщение). Наличие полезной избыточности в языке и речи делает этот обмен аналогичным передаче информации с помощью корректирующих

кодов по методу «кодовой книги» (то есть по максимуму сходства). О свойствах языка как помехоустойчивого кода не раз писали основоположники кибернетики, они подтверждаются также при анализе данных психоакустики [7].

Таким образом, альтернативой СЛ является «мягко-системный» или иначе информационно-стохастический (ИС) подход. Сущность его состоит в том, чтобы выявить и алгоритмизировать механизмы речемыслительной деятельности человека, которые как бы стоят над закрепленными узусом синтаксическими и семантическими правилами.

2.3. Традиционные модели

2.3.1. Концепция семантических падежей Филлмора

Одной из первых лингвистических моделей, получивших широкое распространение в области обработки языка, стала концепция *падежных фреймов*. Первоначальная концепция, ориентированная на проблемы теоретической лингвистики была предложена Филлмором [8] и предусматривала небольшой набор универсально применимых падежей. Теория связана с определением ядерных грамматических отношений в предложении (субъекта, объекта и косвенного объекта) в том смысле, что она задается вопросом: каким образом конкретные аспекты значения высказывания определяют, какая составляющая выступает в качестве (глубинного) субъекта, а какая — в качестве объекта.

Одним из важнейших элементов теории глубинных падежей является *падежная рамка* (frame), которая приписывает семантико-синтаксические роли конкретным участникам (реальной или воображаемой) ситуации, коррелированной с предложением. Филлмор выделил следующие роли (=семантические падежи): агент — одушевленный инициатор действия; контрагент — сила, против которой направлено действие; объект — вещь, являющаяся объектом действия; адресат — лицо, в пользу (вред) которого совершается действие; пациент — вещь которая испытывает эффект действия; результат — вещь, которая возникает в результате действия; инструмент — стимул, или непосредственная физическая причина действия; источник — исходное состояние объекта до выполнения действия [9].

Теория семантических ролей Филлмора послужила источником большого числа работ в области моделирования языка. Последователями данной теории стали разработчики теории концептуальных зависимостей, о которой говорить в следующем разделе.

2.3.2. Теория концептуальных зависимостей

Теория концептуальных зависимостей ставила своей целью обеспечить «понимание» текста естественного языка. Для этого в начале семидесятых, группой Шенка [10], была предложена система смыслового представления, которая должна состоять из понятий и отношений между понятиями. А также должны быть введены ограничения на то, что понимать под тем и другим. Была разработана модель, называемая моделью концептуальных зависимостей (КЗ). Концептуальное действие в ней определяется как нечто, что может быть сделано *актором с объектом*. Выделяются две категории действий: физические

действия (действия, выполняемые над физическими объектами) и мыслительные действия (манипуляции, выполняемые с идеями и т.п.).

Значительное внимание модель КЗ уделяет выводу (=умозаключениям). Умозаключения являются одним из основных средств восполнения смысловой неэксплицитности в модели КЗ. Под умозаключением понимается логическая операция над семантическим представлением текста, полученным после этапа анализа. Умозаключения выполняются с учетом данных об окружающем мире и множества аксиом о человеческом поведении.

С точки зрения высокоуровневой обработки языка и речи модель КЗ является моделью языка и, по существу, занимается исчислением смысла высказывания через так называемые «атомы смысла», в ущерб модели общения. Говоря о парадигме «атомов смысла», можно сказать, что и в Древнем Мире (Платон, Сократ), и в Средние века (Лейбниц), и многие в XIX — XX веках безуспешно пытались вычислять смысл текста из неких «атомов смысла». Однако приемлемых результатов до сих пор не получено.

МКЗ, как и другие теории подобного типа, по существу оперируют неточными исходными данными с помощью точных правил. Это противоречие препятствует успешному применению этого метода.

2.3.3. Сетевые модели

Сетевое моделирование изначально использовалось для построения модели текста, модель включает три основных операции: выявление семантических связей между элементами текста, их представление в удобной для дальнейшего исследования форме и установление (с использованием этого представления) закономерностей, характеризующих изучаемый объект. Семантическая связь между единицами текста отражает связь между соответствующими денотатами. Если денотатом слова является предмет, то денотатом предложения — ситуация (конечное множество предметов, соединенных отношениями). В зависимости от того, какие именно единицы текста приняты в качестве основных элементов системы, различают два вида речевой семантической сети:

- 1) Пословная речевая семантическая сеть — это сеть, вершины которой соответствуют номинативным лексическим единицам, а ребра — обозначают синтагматические отношения.
- 2) Пофразная речевая семантическая сеть — это сеть, вершины которой соответствуют предложениям (точнее, смыслам предложений), а ребра — семантическим отношениям между ними. Для характеристики семантических связей между предложениями основную роль играет констатация наличия связи и ее силы.

В качестве сетевых методов семантического анализа текстов, так называемого семантического представления текста, членения, выделения фрагментов с наибольшей смысловой нагрузкой, семантического вывода, смысловой компрессии текста и т.д., были выработаны способы количественной оценки плана содержания текста, экспликация свойств текста, характеризующих его как систему семантических единиц и отношений между ними, создана методика анализа (классификация, членения, преобразования и т.д.) семантических структур текста [11].

В области речевых технологий потребовалось дальнейшее исследование сетевых моделей. Чтобы учесть существующую в задачах распознавания речи

проблему интеграции разнородных знаний [12] потребовалось создание сетей, объединяющих разнородные знания. В системе распознавания речи, описанной Редди [13], разработан и применен метод, в котором лексические, синтаксические, семантические и прагматические знания объединены в сеть, позволяющую производить анализ любого произношения с учетом всех этих видов знаний. Входной сигнал иерархически преобразуется в сегментированную и фонетически обозначенную форму, которая сравнивается с узлами сети с целью получения оптимального пути по сети. В интегральной сетевой модели используется метод поиска вдоль графа, согласно которому все узлы, кроме совокупности альтернатив, «наиболее близких» к оптимальным, удаляются из поискового дерева в каждой точке ветвления.

2.3.4. Представление знаний во фреймах

Согласно М. Минскому, фрейм (frame) есть воплощение «структуры данных» для выражения некоторой стереотипной ситуации; он представляет собой совокупность вопросов, которые можно задать о соответствующей воображаемой ситуации. Содержанием слотов являются фактические ответы, которые определяют, какие ситуационные аспекты и с помощью каких средств подлежат рассмотрению [14].

Фреймовая теория широко используется в различных теориях понимания речи. Например, представления на семантических сетях сводятся к фреймоподобным массивам ячеек, фреймы используются для представления знаний в сетях информационного поиска. Теория надежных грамматик также базируется на фреймовых структурах [15].

Идея ситуационных рамок — фреймов как способов организации «знаний» машины о некоторой (стандартизованной) внешней среде получила развитие в ряде работ американских кибернетиков. В одной из них — работе Р. Шенка и Р. Абельсона [16] фрейм представлен как структура сведений о реальности, облегчающая обращение машины к этим сведениям; фреймы подразделяются на статические — сценарии (scripts) и динамические — планы (plans). Сценарии и планы служат отображению причинно-следственных связей между ситуациями предметной области, таким образом, фреймовая теория может отразить сюжетные линии и сценарии теории концептуальных зависимостей.

Из вышеизложенного следует, что теория фреймов — это абстрактная теория описаний сети, терминалов, предварительных критериев, преобразований, заранее заготовленных значений, ожиданий и поиска информации.

2.3.5. Некоторые математические аспекты высокоуровневой обработки текста

Целесообразность переноса свойств живого интеллекта на системы искусственного интеллекта (в том числе на системы понимания речи) сейчас почти ни у кого не вызывает возражений.

Сложность математического моделирования процессов восприятия (слухового, зрительного и др.) кроется в недостаточной изученности механизмов мозга. Слишком мало известно, по каким законам происходит накопление сенсорной информации, какие преобразования претерпевает эта информация в ходе формирования признаков, данных, представлений, понятий и знаний, по

каким законам объединяются (взаимодействуют) входные и априорные данные в процессе узнавания, интерпретации и выработки решений о целесообразном поведении.

Помимо незнания функциональных закономерностей дело усложняется из-за стохастической природы языка, которая стала признаваться сравнительно недавно. В противовес начальным представлениям о логически четкой, правильной структуре и детерминированном характере обработки речевой информации в центральной нервной системе, сейчас в результате ряда исследований выясняется, что мозг в большинстве случаев получает нечеткую информацию (недетерминированную, неоднозначную, неполную, неточную), причем компоненты этой информации избыточны и взаимозависимы. Эти кажущиеся недостатки на самом деле оборачиваются рядом важных достоинств, которые отсутствуют в строгой, детерминированной, неизбыточной системе знаний — высокая надежность, гибкость, устойчивость, способность к самокоррекции, но изучение этих закономерностей путем моделирования остается очень сложным.

В соответствии с изложенным, методы математической обработки таких данных и знаний в основном должны быть нацелены на минимизацию риска принятия решения о распознавании или интерпретации и носят в той или иной форме вероятностный, или, во всяком случае, квантитативный характер. При этом математический аппарат для оперирования нечеткими данными на разных уровнях отображения развит в различной степени. Случайность, размытость параметров речевого сигнала на нижнем, акустико-фонетическом уровне была ясно осознана многими исследователями уже в 60-х годах XX века. Ярким выражением этой стадии развития теории явилась пионерская работа Т.К. Винцюка [17], в которой он сумел объединить методы статистической обработки сигнала с новейшим методом оптимизации на основе только появившегося тогда математического метода динамического программирования [18].

Далее в обработке речевого сигнала нашли применение и развитие методы марковского моделирования, градиентные методы. Все успешно работающие приемы такого рода в какой-то мере являются разновидностями вариационного исчисления, приспособленного к современной компьютерной информационной технологии.

Для оперирования нечеткими знаниями высокого уровня сейчас все более используются вероятностные модели языка. Появилось также много работ, суть которых хорошо сформулировал Форсайт [19] как «...создание методов, позволяющих быть точными в отношении неточностей». Среди таких подходов, прежде всего, следует назвать теорию размытых множеств Л.Заде [20]. К сожалению, данный метод пока не нашел широкого применения в языковых задачах, поскольку он базируется на очень сильных упрощениях лингвистических проблем, а также не учитывает статистической природы языка и речи. Далее можно упомянуть ряд оригинальных подходов к языковым задачам: метод коэффициентов уверенности Шортлиффа [21], метод шансов Харта Нильсона [22], функция доверия Демстера-Шафера и вероятностная логика Нильса [23]. Все эти методы пока не нашли широкого применения в обработке языка и речи.

2.4. Альтернативные подходы

2.4.1. От правил к количественным оценкам

Мы рассмотрели несколько традиционных подходов, основанных на принципах математической лингвистики, которые применялись для высокоуровневой обработки текста для таких приложений как реферирование и индексирование текста, создание автоматизированных систем лингвистического информационного обеспечения и др. К сожалению, такие методы, и в частности, многочисленные варианты грамматического и синтаксического разбора не удается успешно применить в области обработки речи. Как заметил известный руководитель ряда международных конференций в области распознавания речи А. Замполли [24]: «За 40 лет вычислительной лингвистики мы не получили программ семантического разбора способных анализировать реальный текст; большинство таких программ способно анализировать лишь несколько высказываний, и затем сбиваются». Результаты международных конференций последних лет в области понимания речи показали, что на основе вышеупомянутых методов можно достичь уровня понимания разговорной речи не выше 75% [25–27], т.е. известные методы обработки текста не дают желаемых результатов для понимания речи. Они должны быть радикально переработаны или заменены принципиально новыми методами.

В противовес методам, основанным на логических принципах, сейчас наблюдается бурное развитие стохастических методов. Это стохастические грамматики, стохастические модели разбора, статистические модели языка и др. Они дают более высокие результаты и видимо лучше соответствуют природе человеческого восприятия языка и речи.

Стохастические грамматики стали развиваться сравнительно недавно в противовес обычным грамматикам строгих правил. Одна из первых работ в этой области [28] посвящена новому механизму для вывода правил стохастической контекстно-независимой грамматики из массива тренировочных данных. Подход основан на теории Перла оценки степени доверия в причинно-следственных деревьях [29]. Вероятностные оценки, получаемые с помощью грамматических правил используются для построения деревьев разбора по сегментам высказывания. Затем применяется оценка доверия, чтобы получить распределение постериорной вероятности для каждой грамматической категории. Этот алгоритм не требует каких-либо меток или сегментации входного сигнала и, следовательно, является полностью безнадзорным (обучение без учителя). Этот метод, по мнению его автора, дает значительно лучшие результаты, чем широко известный метод биграмм и триграмм.

В статье Вольфганга Минкера [30] представлено исследование LIMSI — CNRS в области выявления стохастического компонента понимания естественного языка (Natural language) и приложения стохастического компонента к различным задачам и языкам. Тестирования этого компонента проводились в американском приложении ATIS (Air Travel Information Services) и французском MASK (Multimodal-Multimedia Automated Service Kiosk). Исследования показали, что для ограниченных приложений стохастический метод вырабатывает хорошо настроенный и основанный на правилах компонент. Продемонстрировано, что с помощью такого стохастического компонента человеческие усилия по отнесению данных к определенной категории могут быть сокращены, и это гораздо проще, чем оформление, поддержка и расширение грамматических правил. По-

сколькx стохастический метод автоматически учится семантическому формализму, он является сравнительно легким и гибким методом.

2.4.2. Теория речевых актов

Одной из интереснейших лингвистических теорий, которая внесла весомый вклад в исследование речевых процессов, является «Теория речевых актов» ТРА, предложенная Дж. Остином, и развитая Дж. Серлем и другими [2].

Объектом исследования в ТРА является акт речи, который произносится в ситуации непосредственного общения со слушающим с учетом широкого ситуативного контекста. Таким образом, ТРА способствует наиболее адекватному пониманию сути диалога и принципиально важна для построения систем диалога.

В отличие от других моделей общения, наряду с такими компонентами, как говорящий, слушающий, высказывание, обстоятельства модель речевого акта в ТРА включает в себя также цель и результат речевого акта. Речевой акт рассматривается, как намерение достичь конкретной желаемой цели.

Предварительно в теории выделяются перформативные высказывания, представляющие собой не истинное или ложное сообщение, не только, и не столько говорение, сколько действие. Таким образом, теория определяет главный объект коммуникации и пытается классифицировать перформативные высказывания по типам действий.

В ходе развития теории классификация перформативных высказываний по типам действий претерпевала изменения от классификации перформативных глаголов до классификации речевых актов. ТРА развивалась в соответствии с современными представлениями о важности учета ситуативного контекста для понимания устной и письменной речи. ТРА выявила сопутствие внеречевых актов речевым, т.е. указала на то, что речевые акты имеют коммуникативное намерение. Первый разработчик этой теории, Дж. Остин, подчеркнул, что основной единицей речевой коммуникации является не слово или другая лингвистическая единица, а речевое намерение говорящего, и его распознавание является основной целью, как в общении людей, так и в задачах человеко-машинного диалога.

2.4.3. Использование ситуативного контекста в высокоуровневой обработке речи

Использование естественного языка в общении всегда диктуется целью говорящего. Эта цель сама по себе не предполагает точной формулировки, но накладывает скрытые ориентиры, которые помогают ограничить неоднозначность высказываний в меру необходимости решения соответствующей задачи. Наша способность использовать глобальный контекст для требуемого уменьшения неоднозначности, не прибегая к явной формализации, связана с периферийным сознанием. Это краевое сознание учитывает скрытые ориентиры, заключенные в контексте, а также некоторые грамматические конструкции и в конечном итоге все это должно быть четко сформулировано для машины [31].

Произнесение и понимание предложения естественного языка предполагают достаточное знание того, как предложение связано с контекстом. Поэтому, для того чтобы заставить машину понимать и решать задачи речевого взаимо-

действия (такие как перевод, речевое управление и т.д.) необходимо научить машину связывать слова и словосочетания с ситуациями реальной действительности. Существует два подхода к решению этой задачи. Один подход основан на скрупулезном описании предметной области, т.е. универсальном методе представления знаний. Например: Стул — это мебель, служит для сидения и т.д. Такое описание громоздко и вывод умозаключений на основе таких знаний сложен. Второй подход непосредственно вытекает из ТРА, рассмотренной в разделе 2.4.2, он основан на ситуативной модели деятельности и является предметом исследований СПИИРАН.

2.4.4. Проблема интеграции знаний

Одними из первых Шенк и Бирнбаум [12] предложили идею интегральной обработки, по которой понимание языка и речи достигается объединенным применением синтаксических, семантических и прагматических знаний. Развивая эту гипотезу, Селфридж [22] сформулировал понятие надежности (robustness) для естественных языковых систем, которая возникает вследствие использования интегрального подхода. *Надежная понимающая система — это такая система, для которой мы можем гарантировать, что она в итоге понимает входные высказывания, несмотря на пропуски произвольных слов, на любые нарушения порядка слов и на наличие эллипсиса, — как с использованием, так и без использования поправок; при этом она опирается на семантику и синтаксис, знания о предметной области и контексте.* Для создания надежной (robust) системы понимания Селфридж считал необходимым определить отношения между синтаксисом и семантикой, а также между пониманием языка и обработкой памяти, для этого он опирается на вопросы, сформулированные Шенком и Бирнбаумом:

- 1) Обработывается ли синтаксическая структура раньше семантической или синтаксис и семантика обрабатываются одновременно?
- 2) Разделены ли процессы обработки синтаксиса и семантики или обработка этих структур образует единый процесс?
- 3) Являются ли обработка языковых сообщений и обработка знаний, содержащихся в памяти, разными процессами или они составляют принципиально один и тот же процесс?

Как отмечают Шенк и Бирнбаум, существуют, грубо говоря, две полярные точки зрения на эти вопросы. Одну из них можно назвать «сепаратистской», а другую — «интегральной». Каждую из них можно охарактеризовать с помощью ответов на указанные вопросы. Первый вопрос касается временного соотношения семантической и синтаксической обработки в ходе процесса понимания. Согласно сепаратистской точке зрения, синтаксический анализ высказывания осуществляется до любого вида семантического анализа, а на его выходе мы получаем синтаксическое описание высказывания. Это описание затем подается на вход семантического анализа. Т.о. производится последовательный анализ высказывания, при котором возможно верная гипотеза может быть отвергнута на одном из уровней предварительного отбора. Такому представлению противостоит *интегральная концепция, согласно которой синтаксический анализ выполняется в то же самое время, что и семантический анализ.*

Второй вопрос касается природы механизмов, обрабатывающих синтаксические и семантические структуры. В соответствии с сепаратистской точкой зрения механизм, строящий синтаксическое описание высказывания, отличен

от механизма, строящего представление значения высказывания, т.е. семантический анализ осуществляется по иному алгоритму и поэтому их необходимо выполнять раздельно. *Интегральная точка зрения, напротив, утверждает, что синтаксис и семантика обрабатываются одним и тем же механизмом.* Этот механизм одинаково успешно оперирует как синтаксической, так и семантической информацией.

Третий вопрос касается взаимоотношений между обработкой языковых сообщений и обработкой памяти. Сепаратистская позиция состоит в том, что обработка языка — это особая, специфическая функция, во многом не связанная с процессами обработки памяти. В рамках этого взгляда память трактуется как относительно пассивная сущность, лишь в малой степени затрагиваемая активной обработкой. Интегральная точка зрения состоит в том, что обработка языковых сообщений — это, прежде всего процесс, основанный на операциях с памятью, следовательно, обработка языка и обработка памяти должны составлять единый процесс.

Гипотеза об интегральной обработке правильно отражает природу обработки языковых сообщений человеком, что ведет к предположению, что интегральная обработка обеспечивает надежное понимание.

Как видно из [12, 22] данная гипотеза недостаточно развита. Нет конструктивного определения «интегральной обработки». Другое определение этого понятия предложено нами в разделе 4. Принцип «одновременности», предложенный вышеуказанными авторами, ограничивает применение данной идеи. Например, акустико-лексическая обработка заведомо предшествует семантико-синтаксической обработке, однако, интеграция этих уровней желательна, возможна, что и показано в разделе 4.

2.5. Заключение

В заключение данной главы необходимо сказать следующее.

- 1) В рамках исследований методов обработки текстов было создано много подходов обработки языка. Однако ввиду существенных различий речи и текста эти методы оказались в большинстве своем непригодны для эффективной обработки речи.
- 2) Практика исследований речи показала хорошую перспективу стохастических и количественных методов.
- 3) Наиболее перспективными направлениями обработки языка следует считать идеи, которые хорошо коррелируются со знаниями о природе человеческого восприятия (количественные методы обработки, интеграция разнородных знаний, учет ситуативного контекста, ассоциативные процессы, феномен сквозного понимания и др.)

3. Состояние исследований автоматического понимания речи

Согласно общепринятому в мировой практике определению, понимание речи состоит из двух последовательных этапов: распознавание речи (SR) и понимание естественного языка (NLU). Как будет показано далее, такой разобщенный подход не является оптимальным, поскольку он игнорирует концепцию интегральной обработки. В этом разделе кратко представлены типовая структура системы взаимодействия человек-машина, в рамках которой компьютер

осуществляет «понимание» фраз; состояние проблемы распознавания речи; состояние проблемы понимания речи в различных диалоговых системах; дается оценка существующим подходам к методам понимания речи.

3.1. Типовая структура системы взаимодействия человек-машина

Типовая структура системы взаимодействия человек-машина в самом общем виде показана на рис.2. Процесс диалога обычно инициируется речевыми актами. Человек принимает решение о том или ином акте на основе своих априорных знаний о теме и цели диалога, о языке диалога, а также всей доступной текущей информации о ходе диалога, внешней среде и текущем времени.

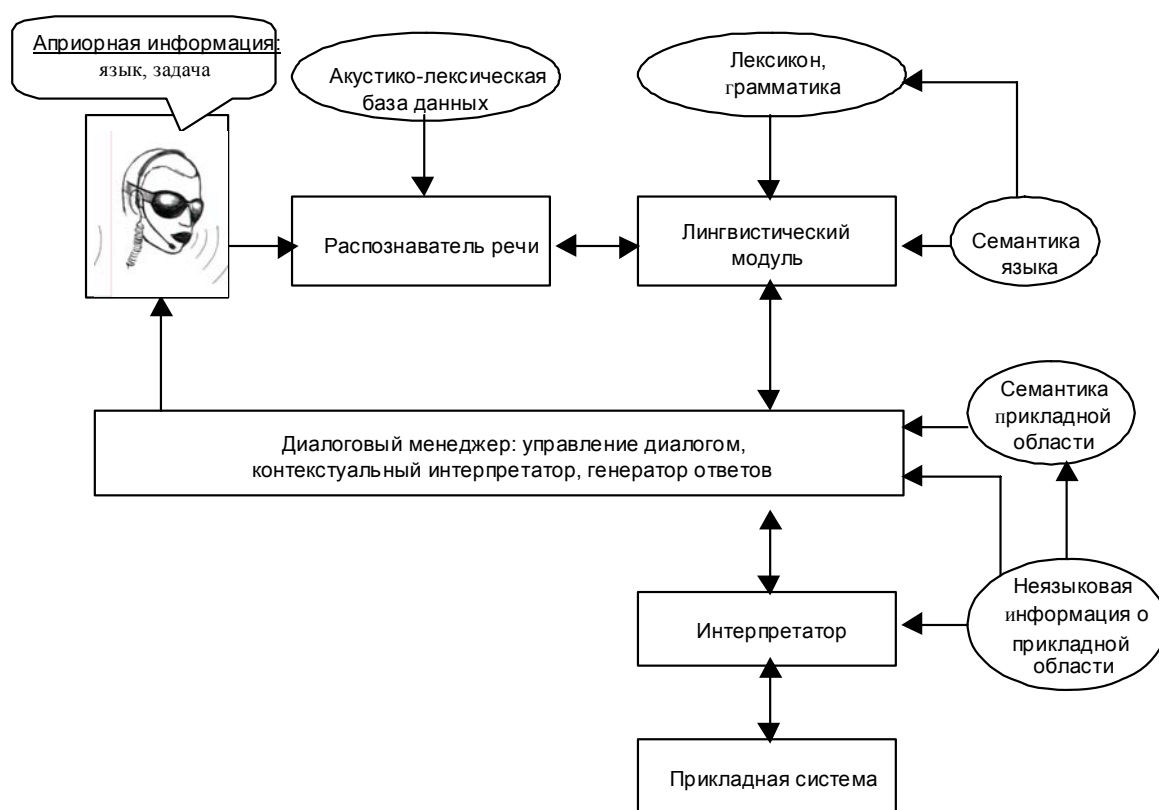


Рис. 2. Типовая структура диалоговой системы

Модуль распознавания речи распознает акустический речевой сигнал, т.е. выдает соответствующую последовательность слов. Эта последовательность слов поступает в подсистему понимания языка, которая вырабатывает смысловое представление фразы в виде семантического фрейма, в виде указателя типа смысла и др.

Система понимания использует знания о языке предметной области и текущей ситуации.

Координацию всех основных звеньев диалоговой системы осуществляет система управления диалогом (менеджер). В простейших системах этот модуль отсутствовал, но с усложнением процесса диалога появилась необходимость,

например, уточнения некоторых слов, подсказки оператору, сведения об ограничениях и т.д., и таким образом, менеджер является неотъемлемой частью современных диалоговых систем.

Надо заметить, что в простейшем случае вначале обходились без систем понимания, а использовали распознавание слов и систему меню [42,84], но этот метод вряд ли можно назвать интеллектуальным. Менеджер также либо не использовался, либо его функции сводились к подтверждению пользователем правильности распознавания слова. Естественно, эффективность таких систем не высока.

В настоящее время проблема управления диалогом приобрела большое значение в связи с тем, что задачи диалога быстро развиваются и усложняются. Появились исследования вопросов стратегии диалога [32,33], структуры диалога [34,35], обучения «менеджера» [36,37,38] и др.

3.2. Методы распознавания речи

3.2.1. Основные проблемы распознавания речи

Процесс распознавания речи представляет собой преобразование акустического сигнала, полученного от микрофона, в последовательность слов. Полученный набор гипотез цепочек слов далее используется для понимания речи.

Задачи распознавания речи характеризуются многими параметрами, в первую очередь, это уровень окружающего шума, свойства канала передачи речи, размер словаря, вариативность речи, тип ввода речи (изолированная/слитная). Для распознавания изолированных слов необходимо, чтобы диктор делал короткие паузы между словами, что замедляет ввод и ухудшает естественность, в то время как при вводе слитной речи этого не требуется.

Сложность проблемы распознавания речи, главным образом, связана с вариативностью ее основных параметров, на которые влияет множество факторов. Прежде всего, это случайная компонента процесса речеобразования, которая приводит к многообразию описания одного и того же слова, произнесенного одним и тем же диктором. Более существенный фактор — это индивидуальное различие речевых аппаратов разных дикторов. Далее надо отметить влияние пола диктора, возрастных различий, диалектов, эмоционального и физического состояния диктора. Значительное влияние вносит акустический аспект, т.е. смена микрофона, расположение микрофона относительно рта, акустическая обстановка в помещении.

Для решения проблемы вариативности речи на уровне параметрического представления сигнала стараются выбрать дикторонезависимые параметры. Кроме того, возможно использование методов адаптации, позволяющие настроить систему на конкретный голос и акустическую обстановку.

Мерой качества акустико-лексического распознавания слов может служить точность распознавания некоторого фиксированного словаря. Возможности этого уровня объективно ограничены вышеизложенными факторами и даже для человека могут составлять, например, точность распознавания всего 85% на словаре в 300 слов [39]; в этом эксперименте были исключены синтаксические и семантические связи, чтобы вынудить человека распознавать слова только на основании акустическо-лексической информации.

Распознавание слов существенно ухудшается с увеличением словаря, так как при этом, появляются группы акустически подобных слов, что приводит к путанице, причем эта путаница резко усиливается с ростом словаря. Есть два пути преодоления этой трудности. В первом случае используют достаточно большой запас гипотез, такой, чтобы фактически введенные слова оказались в составе этих гипотез, тогда остается принципиальная возможность правильного распознавания смысла в некоторой системе понимания. Однако с ростом словаря этот запас должен резко возрастать и для 50000 слов может составить тысячи гипотез на каждое входное слово. Другой подход основан на отборе в ходе распознавания речи лучших гипотез слов с использованием некоторой высокоуровневой информации. Для этого стали привлекать высокоуровневую информацию, обычно синтаксическую и семантическую в виде стохастических моделей языка. Одной из первых работ в этой области была модель распознавания 5000 слов фирмы IBM на основе сочетания акустико-лексической обработки с моделью языка, основанной на статистике биграмм и триграмм слов. Точность распознавания составляла 95% [40]. Такая парадигма сохраняется и до наших дней, основным направлением в акустической обработке является глубоко развитая техника скрытого марковского моделирования (СММ), которая обеспечивает дикторонезависимость и способность распознавания слитной речи. Сейчас имеется целая серия аналогичных промышленных систем для словарей с объемом до 60000 слов и более [41].

Хочется подчеркнуть, что такие модели изначально создавались для систем автоматической стенографии, в которых предполагалась последующая ручная корректировка. Сейчас наблюдаются многочисленные попытки применить эти системы для различных интеллектуальных приложений, таких как справочные системы [42], голосовое управление [43], системы перевода [44, 45] и др. Однако эти попытки не достигают желаемых результатов, что по нашему мнению происходит из-за некорректного использования высокоуровневой информации. Дело в том, что в таких моделях отсутствует уровень понимания, а наблюдения показывают, что эти процессы присутствуют в работе человека-переводчика, стенографистки, редактора и т.д. Другой недостаток состоит в том, что на уровне распознавания слов интегрально используется два вида знаний: акустико-лексические и семантико-синтаксические в виде стохастической модели языка, что может приводить к последующему конфликту знаний на этапе понимания речи. Выход из этого кризиса видится только в процессе тесной интеграции знаний на этапе понимания, о чем будет подробнее сказано в разделе 4.

3.2.2 Методы параметрического представления сигнала

При построении системы распознавания речи одной из основных задач является выбор системы параметров, которая позволила бы достаточно хорошо различать звуки и слова речи и в то же время оставаться инвариантной к особенностям произношения конкретного диктора, к изменениям акустической обстановки, смене микрофона и т.п.

Существует широкий спектр методов параметрического представления сигнала. Первые системы распознавания речи в качестве предварительной обработки сигнала использовали автокорреляционный анализ, аппаратную полосовую фильтрацию, расчетный спектр, метод коэффициентов линейного предсказания (КЛП). Сейчас наиболее распространены два класса методов, использующих разновидности спектрального анализа и КЛП. Популярность этих ме-

тодов обусловлена тем, что они хорошо согласуются соответственно с моделями слухового восприятия и речеобразования человека.

Речевой сигнал, как правило, записывается с частотой дискретизации 8 — 20 кГц. Эта последовательность цифровых отсчетов разделяется на сегменты речи длительностью 10 — 20 мс. Для каждого такого сегмента вычисляется вектор параметров, эти векторы составляют основу параметрического описания речи. Такой анализ называется кратковременным.

Набор полосовых фильтров является одним из фундаментальных понятий в распознавании речи. Набор фильтров соответствует модели слухового аппарата человека. Спектральное описание сигнала, получаемое с помощью системы фильтров можно рассматривать как реакцию механической системы внутреннего и среднего уха на воздействие сложного сигнала. При этом различные составляющие этих колебаний заставляют вибрировать отдельные участки базилярной мембраны [46]. Поэтому формирование набора полосовых фильтров, производится с учетом основных свойств слухового аппарата человека, который наилучшим образом приспособлен к задаче распознавания речи.

Изучение спектрального преобразования сигнала привели многих исследователей к весьма сходным алгоритмам вычисления оптимального набора параметров. Эти алгоритмы включают в себя предварительную обработку цифровых отсчетов, например предсказывающий фильтр и процедура весового окна [47,48], быстрое преобразование Фурье (БПФ) над сегментами речи 10-20 мс, формирование набора перекрывающихся фильтров, расположенных эквидистантно или по некоторому нелинейному закону, например, согласно *Mel* [49] или *Bark* шкале [50]. Отсчеты БПФ, включенные в каждый фильтр, пересчитываются с учетом треугольного окна, затем определяется интегральная энергия в каждом фильтре и далее производится логарифмирование выхода каждого фильтра. Этот набор данных подвергается косинус преобразованию, что в итоге приводит к так называемым *кепстральным* коэффициентам [51,52]. Кепстральные коэффициенты, полученные с использованием *Mel*-шкалы, называются мел-частотными кепстральными коэффициентами.

Техника линейного предсказания [53], основанная на авторегрессионном анализе, сегодня менее популярна в распознавании речи, но все еще широко распространена в системах сжатия речи. Основной принцип линейного предсказания основан на аппроксимации текущего отсчета речевого сигнала через линейную комбинацию соседних отсчетов. Коэффициенты предсказания — это весовые коэффициенты, используемые в линейной комбинации, которые вычисляются путем минимизации среднего квадрата разности между отсчетами речевого сигнала и их предсказанными значениями. Алгоритм вычисления вектора признаков с помощью КЛП представлен на рис. 3. В итоге мы получаем описание, адекватное модели речевого тракта. Более полное описание речи может быть получено путем объединения информации о речевом тракте и источнике возбуждения.

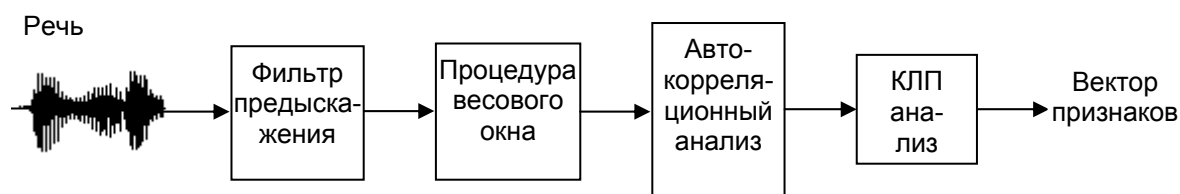


Рис. 3. Типовая модель обработки сегмента речи с помощью КЛП

Поскольку для хранения параметрического представления речи требуется довольно большой объем памяти и, учитывая значительную случайную составляющую речи, целесообразно сжать эту информацию. Для этой цели применяется техника векторного квантования [54]. Наиболее типичная схема использования этой техники представлена на рис. 4. Основная идея этой техники состоит в том, чтобы отобразить неограниченное множество векторов признаков речевого сигнала на конечное множество наиболее типичных акустических состояний. При этом многомерный вектор параметров заменяется индексом соответствующего элемента кодовой книги. Кроме того, закрепляя за элементом кодовой книги определенный звук, мы устанавливаем связь с фонетическим уровнем, что важно для последующей обработки.

Важным вопросом векторного квантования является разработка соответствующей кодовой книги, необходимой для квантования, а также выбор меры расстояния между векторами и процедуры выбора ближайшего элемента. При этом используют сравнение со всеми элементами кодовой книги, или иерархические процедуры кластеризации [55, 56].

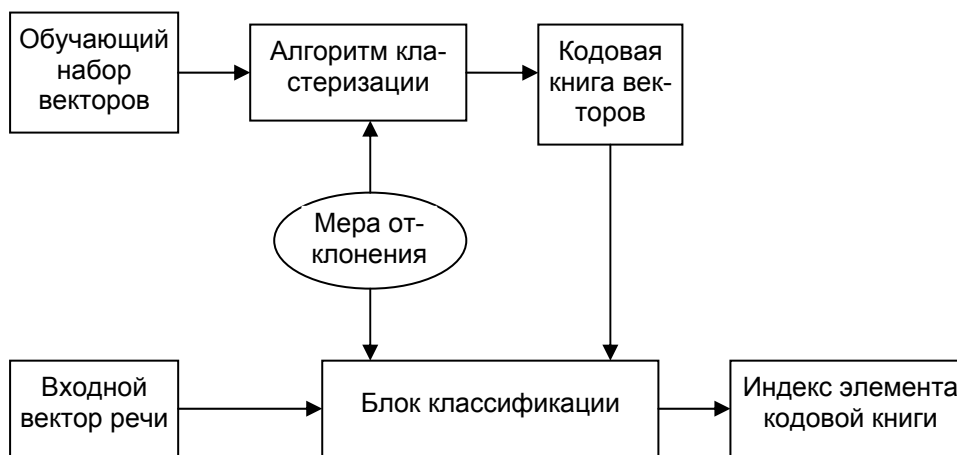


Рис.4. Блок схема обучения кодовой книги и классификации векторов

Процесс построения кодовой книги известен также как процесс обучения или заполнения кодовой книги. Наиболее популярны сейчас методы K-средних [57] и ISODATA [58].

3.2.3. Распознавание речи с помощью динамического программирования

Ключевым вопросом в распознавании речи является идентификация элементов речи, которая ведется или путем сравнения с эталонами (фонем, слов, словосочетаний), или посредством оценки правдоподобия принадлежности данного сигнала к тому или иному классу.

Наиболее простой задачей является распознавание изолированных слов. Для этого необходимо предварительно определить границы слов в контексте окружающих шумов. Сложность определения границ слов связана, во-первых, с особенностями произношения конкретного диктора, с наличием в его речи шумов процесса артикуляции (придыхание, чмоканье и т.п.). Кроме того, важен учет посторонних шумов. Совершенно очевидно, что точность распознавания

изолированных слов существенно зависит от точности определения их границ. Блок схема итеративного алгоритма определения границ речи приведена на рис. 5.

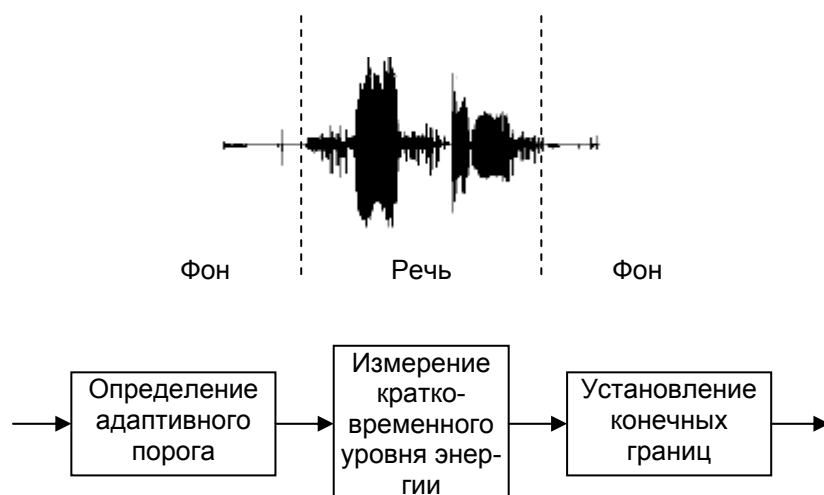


Рис.5. Базовый алгоритм определения границ речевого сигнала

Основная трудность сравнения отрезков речи связана с вариативностью временного масштаба речи, с его нелинейными флюктуациями, поэтому алгоритмы сравнения тесно связаны с задачей нейтрализации деформаций темпа. Изменчивость темпа речи выражается в неконтролируемых флюктуациях продолжительности звуков речи, их участков и пауз. Для того чтобы сравнить слово с эталоном, надо путем деформации оси времени совместить участки, соответствующие одним и тем же звукам, измерить остаточные различия между ними и просуммировать эти частные расстояния, взятые с некоторыми весовыми коэффициентами.

Для нелинейного согласования речи наиболее широко использовались алгоритмы динамического программирования (ДП), базирующиеся на фундаментальных работах Беллмана [18]. Одна из первых публикаций по применению ДП в распознавании речи принадлежит Т.Винцоку [17]. Детальное исследование модификаций этого метода было выполнено специалистами японской фирмы NEC [59]. Задача ДП сводится к поиску оптимального согласования двух отрезков речи A и B. Описания слов представлены в виде последовательностей векторов признаков:

$$A = \{a_1, a_2, \dots, a_i, \dots, a_k\};$$

$$B = \{b_1, b_2, \dots, b_j, \dots, b_n\}.$$

В пространстве признаков выбрана некоторая метрика $d(i, j)$, позволяющая определить меру различия соответственно между векторами a_i и b_j . Затем находится оптимальная траектория на фазовой плоскости (i, j) , которая получается в процессе минимизации суммы всех $d(i, j)$, встречающихся на этой траектории, взятых с определенными весовыми коэффициентами. Работа алгоритмов ДП основана на использовании рекуррентных уравнений. Рекуррентное ДП-уравнение с двукратной степенью деформации представлено на рис. 6.

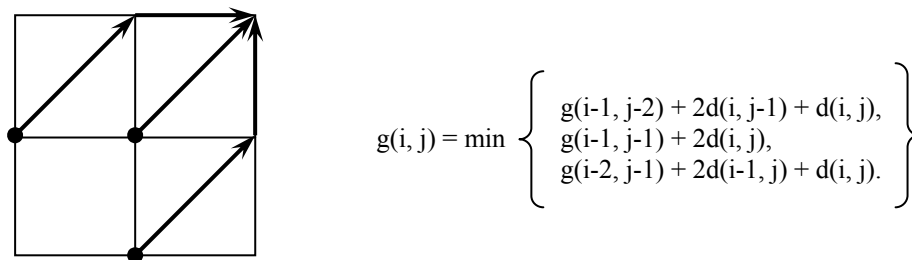


Рис. 6. Рекуррентное ДП-уравнение с двукратной степенью деформации

Разновидности этих уравнений непосредственно связаны с гипотезами о характере взаимных деформаций отрезков речи А и В. Исследование этого вопроса [1,60] показали, что оптимальной деформацией является двукратная, что подтверждается минимумом ошибок при распознавании.

Модель распознавания речи на основе ДП представлена на рис. 7, которая содержит наиболее типичные блоки преобразований: предварительная обработка сигнала с помощью КЛП или спектрального анализа, определение границ необходимого элемента речи, сравнение входной речи со словарем эталонов речи посредством ДП. В результате ДП согласования вычисляется ДП-расстояния, которые позволяют в итоге выбрать оптимальную гипотезу о произнесенном слове.

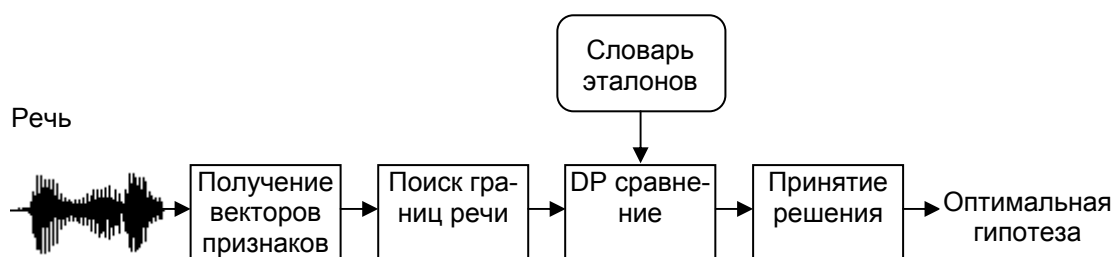


Рис. 7. Распознавание речи с помощью динамического программирования

В последние годы распознавание речи с использованием только метода ДП менее популярно в связи с бурным развитием методов скрытого марковского моделирования. Тем не менее, модификации методов ДП достаточно эффективно решают задачу слитной речи.

3.2.4. Использование скрытого марковского моделирования в распознавании речи

Теоретическую базу для статистического моделирования языковых процессов заложил петербургский профессор А. Марков в начале XX века. Сейчас методами Марковского моделирования пользуются почти все исследователи речи в мире.

Разновидность марковских моделей, называемая *скрытая марковская модель* (СММ), основанная на теории дискретных случайных цепей, была впервые

введена и изучена в конце 60-х — начале 70-х годов [61,62]. По существу, СММ — это дважды стохастический процесс. Термин «дважды» используется для обозначения такой пары процессов, один из которых является основным, но скрытым от нас и наблюдаемым только через другой стохастический процесс [63]. Модели такого типа особенно удобны для описания речевого сигнала, поскольку в действительности давление звуковой волны, которое мы измеряем, представляет собой только некоторый код основного символического процесса, протекающего в ненаблюдаемых и полностью недоступных участках мозга. В наблюдаемом, акустическом процессе выявляются измеримые физические корреляты лингвистической структуры.

Марковские модели довольно содержательны по своей математической структуре, поэтому они стали теоретическим фундаментом для различных областей исследований, не только языка. СММ модель речевого сигнала позволяет решать не только задачи распознавания речи, но также улучшать качество сигнала, загрязненного шумами и искажениями, моделировать источник речевого сигнала, оптимизировать структуру диалога и др.

При построении модели распознавания на основе СММ, выбирают ее основные параметры: тип модели (эргодическая, модель Бэкиса [64] и др.), размер модели (число состояний), тип наблюдаемых параметров.

В качестве единицы распознавания речи можно использовать различные лингвистические единицы, такие как фонема (или фонемоподобные единицы), дифоны, полуслоги и слоги [65], а также производные единицы, такие как фенемы, феноны и акустические сегменты [66]. Каждая подобная единица может быть представлена некоторым типом СММ, настройка параметров которой осуществляется по обучающему множеству речевых данных. На рис. 8. показана структура переходов соответствующей модели Бэкиса для фонемы.

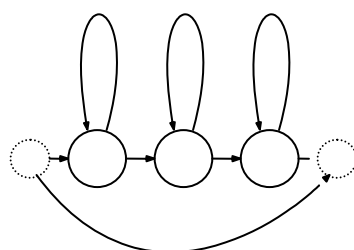


Рис.8. Модель фонетической единицы.

СММ для слова получается путем соединения цепочки моделей фонем, соответствующего алфавита, что представлено на рис. 9.

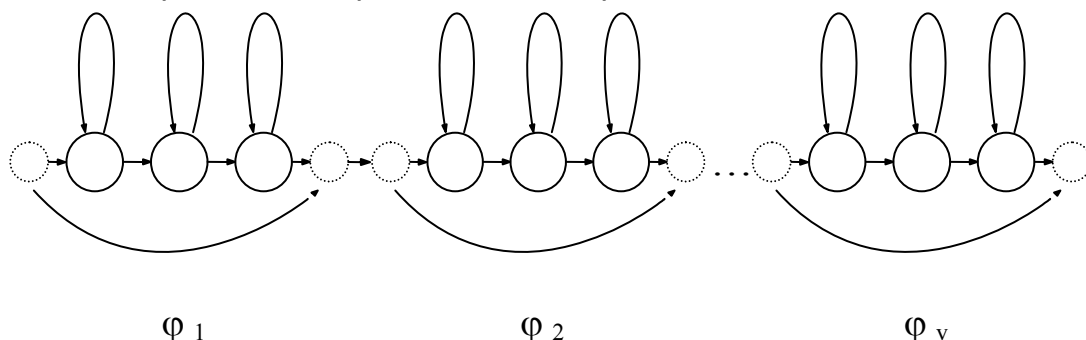


Рис.9. Марковская модель слова, содержащая фонетические элементы из алфавита Φ

В СММ суть распознавания сводится к оценке правдоподобия принадлежности входной последовательности (наблюдаемой) к данной гипотетической модели. Существует несколько способов оценки правдоподобия. Наиболее широко применяется алгоритм Витерби [67], использующий методы ДП, а также алгоритм Баум-Уэлча [62].

Использование марковского моделирования для распознавания изолированных слов можно упрощенно разделить на два этапа: создание СММ для каждого слова из словаря с объемом V , а также оптимизация их параметров и второй этап — распознавание. Для каждого неизвестного слова, подлежащего распознаванию, применяется обработка, показанная на рис. 10, где определяется последовательность наблюдений $O = \{o_1, o_2, \dots, o_n\}$ путем анализа речевого сигнала, затем производится вычисление вероятностей правдоподобия всех возможных гипотез $P(O/\lambda^v)$, где $v \in [1, V]$. Модель, вероятность правдоподобия которой наибольшая, считается оптимальной гипотезой сказанного слова, т.е. индекс распознанного слова v^* вычисляется так:

$$v^* = \underset{v \in [1, V]}{\operatorname{argmax}} [P(O/\lambda^v)].$$

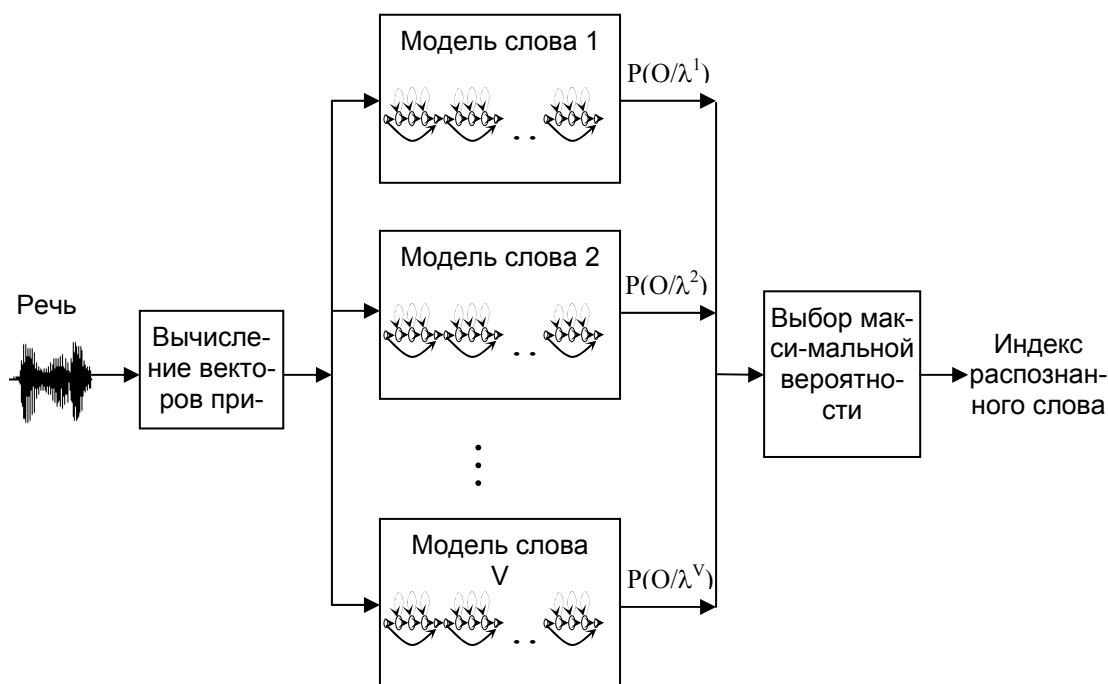


Рис. 10. Распознавание изолированных слов на основе СММ

Разработка систем распознавания, основанных на СММ, требует значительных объемов акустических данных, позволяющих создавать необходимые стохастические базы данных. Сегодня существует ряд популярных акустических корпусов, таких как, TIMIT, RM, ATIS, WSJ и др. При создании речевых баз

учитывается множество факторов, таких как характеристика диктора (национальность, пол, возраст), канал передачи данных (микрофон, телефон), уровень шума. Эти базы данных могут содержать фонетические транскрипции, разметку акустического сигнала по фонемам, слогам, словам, фразам.

Прогресс в развитии систем распознавания речи за счет привлечения аппарата СММ значительно продвинул решение таких проблем, как дикторонезависимость, ввод слитной речи, большие объемы словарей.

3.2.5. Распознавание слитной речи

В отличие от печатного текста или от искусственных сигналов естественная речь не допускает простого и однозначного членения на элементы (фонемы, слова, фразы), поскольку эти элементы не имеют явных физических границ. Они вычленяются в сознании слушателя — носителя данного языка в результате сложного многоуровневого процесса распознавания и понимания речи. Если попросить слушателя записать в виде фонем незнакомую иностранную речь, то он совершит множество ошибок членения слов и фраз, т.е. даже человек не может членить речь без использования знаний лексики, грамматики, смысла.

Границы слов в слитной речи определить невозможно без привлечения всей совокупности априорных знаний о языке, о его конкретном применении. Границы могут быть определены лишь в процессе распознавания, посредством подбора оптимальной последовательности слов, наилучшим образом согласующейся с входным потоком речи по акустическим, лингвистическим и прагматическим критериям. Поскольку этот подбор ведут одновременно с нормализацией темпа речи в условиях неопределенности границ слов, то для его осуществления применяют многоуровневую оптимизацию. Существует три наиболее популярных подхода — двухуровневый алгоритм ДП, метод построения уровней (level-building) и однопроходный (one-pass) метод [51]. Алгоритмы используют одинаковые базовые принципы и отличаются вычислительной сложностью, используемым объемом памяти и сложностью реализации.

Перечисленные методы распознавания слитной речи используются как при сравнении эталонов речи, так и при статистическом моделировании на основе СММ. В последнем случае вместо эталона, представленного последовательностью N векторов, берется СММ с N числом состояний. Задача распознавания речи состоит в подборе оптимальной последовательности моделей слов, которая наиболее вероятна неизвестной последовательности входных слов. В итоге все эти методы опираются на тот или иной способ построения гипотез фраз, используя для этого различные виды генеративных грамматик.

Важным компонентом понимания слитной речи является модель генерации гипотез фраз. Чаще всего используют генеративные грамматики, основанные на строгом синтаксисе. Однако в этом случае модель практически не воспринимает синтаксически неточные, но потенциально понятные фразы, что не позволит в дальнейшем на этапе понимания определить правильный смысл. Самой полной моделью была бы модель генерации всевозможных словосочетаний, как это делается, например, при распознавании слитно произнесенных цифр. Однако очевидно, что при больших словарях число фраз гипотез становится недопустимо большим. Таким образом, здесь требуются соответствующие методы оптимизации.

3.2.6. Распознавание больших словарей

Сложность распознавания слов резко возрастает с увеличением размера словаря; для точного распознавания слов становится недостаточно одной лишь акустическо-лексической информации. Поэтому отличительными особенностями систем распознавания больших словарей (тысячи слов) являются многоуровневая обработка, использование высокоуровневых процедур и знаний. Они реализуются обычно на основе статистических моделей, путем анализа смыслового контекста и с помощью других подходов.

На сегодняшний день созданы системы распознавания в реальном времени слитно произносимых слов из словаря в 60 тыс. слов. Первая публикация о подобных системах была сделана фирмой IBM [40]. В системе использовано марковское моделирование и стохастическая модель языка. В других системах лингвистическая информация реализуется за счет грамматического и контекстуального анализа.

Сочетание информации двух разных уровней в распознавании элементов первого уровня нам представляется не корректным, поскольку выход распознавателя слов будет далее в прикладной системе снова подвергаться высокоуровневой обработке, при этом возможен конфликт знаний. Кроме того, здесь невозможна интеграция всех основных видов знаний, поскольку часть обработки уже сделана «сепаратным» методом. Видимо этим обстоятельством определяется неудачное использование подобных технологий для различных интеллектуальных приложений.

3.3. Методы понимания речи: две основные парадигмы

Задачей понимания в системах диалога является перевод последовательности слов, полученной с выхода распознавателя слов, в некоторое смысловое представление. Сложность задачи понимания связана с большой неопределенностью в распознанной последовательности слов.

Причинами искажений могут быть: ошибки распознавания слов, акустический шум, посторонние звуки, придыхание, чмоканье губами, кашель, озвученные паузы (эээ..., ммм..., и т.п.), повторения, прерывания, пропуски слов, дополнительные слова и т.д. Таким образом, основной задачей понимания речи является нейтрализация вышеперечисленных отклонений.

По общепринятому в зарубежных публикациях определению, задача понимания речи (SLU) состоит из двух задач: распознавание речи (SR) и понимание естественного языка (NLU). Эта последняя задача по-разному решается в рамках двух различных парадигм: *последовательный разбор* и *распознавание количественными методами*.

В рамках первой парадигмы задача NLU решается методами последовательного грамматического и семантического разбора [26, 27, 71] и результатом этого разбора в большинстве случаев является заполненная структура семантического фрейма [72].

К сожалению, системы, построенные на этих принципах, обычно имеют невысокую точность понимания смысла, порядка 70% и не могут быть использованы в коммерческих приложениях. Для преодоления этого недостатка в новейших системах стали применять стохастический компонент в грамматике [73, 74] в сочетании со стохастической моделью языка [75, 76]. Эти системы

содержат некоторую модель разбора и стохастическую модель языка (LM). По данным авторов упомянутых публикаций точность понимания достигает 95%. Дальнейшим усовершенствованием этих методов является добавление информации о конкретной узкой задаче, где кроме обычной модели языка LM используется также ситуативная модель языка LMS [77], которая аналогична обычной LM, но касается конкретной узкой задачи. Такая адаптация системы к задаче, по мнению авторов, повышает устойчивость диалога.

По поводу данной парадигмы надо сказать, что реальная речь насыщена идиомами, выражениями профессионального сленга, всевозможными сокращениями, и она редко позволяет с первого раза заполнить все слоты фрейма, поэтому требуются переспросы, дополнительный ввод слов, что приводит к затягиванию диалога и не соответствует психологии человеческого общения.

Существенным недостатком парадигмы последовательного разбора является то, что семантический разбор нуждается в предварительном точном грамматическом разборе, что породило проблему обнаружения и исправления ошибок [78, 79], которую пытаются решить методами накопления ошибочных фраз. Такого недостатка лишены системы, созданные в рамках второй парадигмы.

В противовес парадигме разбора в настоящее время успешно развиваются методы **распознавания смысла, которые основаны на количественном анализе**, в частности, на стохастической обработке и распознавании смысла путем количественного сравнения цепочки слов с образцами предложений, содержащихся в базе данных [80–82]. Сравнение производится либо на основе стохастических правил [81], либо на основе эвристики [97, 98], либо с использованием разновидностей СММ [83]. В отличие от первой парадигмы такая система способна дать правильный ответ даже при неполной информации, сокращенных фразах, частичных искажениях и т.д. Данный подход согласуется с ранее упомянутой теорией речевых актов. Этот метод, как правило, более оперативен и, по нашему мнению, предпочтителен для задач управления техническими средствами и в других интеллектуальных приложениях.

3.4. Основные проблемы автоматического понимания речи

На основе анализа исследований в области человеко-машинной коммуникации можно выделить следующие ключевые проблемы:

- 1) Главная проблема состоит в создании методов и алгоритмов устойчивого (robust) понимания речи. Здесь хорошим ориентиром могут служить механизмы восприятия человека, которые игнорируют многие виды речевых неточностей и вырабатывают «точный смысл из неточных слов».
- 2) Создание моделей интеграции всех основных видов информации в ходе процесса понимания речи.
- 3) Проблема разумной степени полноты входной информации. Традиционные методы семантического разбора обычно опираются на идею «понимание есть способность ответить на все вопросы, связанные с входным высказыванием». Этот принцип обычно сводится к необходимости скрупулезного заполнения всех слотов семантического фрейма, что ведет к усложнению модуля управления диалогом, к переспросам и к необоснованному замедлению хода диалога. Человек обычно использует интуитивное чувство ситуации и способен понимать различные сокращения, идиомы, профессиональный жаргон. Иными словами, существует некоторый разумный уровень полноты диалоговых фраз, в принципе достаточный для их

надежного понимания без переспросов. Поэтому статистические данные, с достаточной полнотой отражающие речевое поведение человека, могут быть полезными в этой проблеме.

- 4) Усовершенствование модулей распознавания слов в части более быстрой адаптации к голосу, акустической обстановке и конкретной прикладной задаче.

4. Подходы СПИИРАН к проблеме речевого диалога

4.1. Для чего нужно понимание речи?

Этот вопрос не интересовал разработчиков устройств распознавания речи на начальной стадии развития этих исследований, поскольку тогда размер словаря был мал (десятки слов) и все мыслимые приложения обходились однословными командами.

Однако эта проблема давно волнует философов в связи с поиском абсолютной истины. Именно поэтому принцип истинно/ложно в анализе текста дошел до наших дней.

Кроме того, проблема понимания уже в Древнем мире имела огромное значение в связи с судебными, торговыми, политическими и др. спорами, когда возникала необходимость интерпретировать то или иное преступление с точки зрения законов или моральных норм. Как известно, подобные проблемы существуют и в наши дни.

Настоящий бум в понимании речи возник в 50-х гг. в рамках дискуссии «может ли машина мыслить?». При этом возникли разные направления в высокоуровневой обработке текста, а потом и речи о чем говорилось в разделах 2 и 3.

Здесь сразу надо оговориться, что мы не рассматриваем специфические аспекты понимания, такие как извлечение скрытого смысла, решение научных проблем, расследование преступлений и пр. Рассматривается главный аспект понимания — смысловая интерпретация, то есть перевод цепочки слов с выхода распознавателя в смысловое представление. Сложность этого процесса вытекает из уже обсуждавшейся выше проблемы многообразия искажений входной речи за счет внешних факторов, неточностей произношения и ошибок распознавателя.

В связи с появлением распознавателей сравнительно больших словарей возникло желание применить их в различных интеллектуальных приложениях. При этом разработчики всерьез столкнулись с проблемой надежного понимания, которую не удавалось решить за счет традиционных лингвистических подходов. Оказалось, что распознаватели огромных словарей несколько не продвинули проблему речевого диалога. Здесь стоит напомнить об одном курьезном случае. Человек, обладающий феноменальной памятью, заключил пари о том, что он в течение очень короткого времени выучит словарь из 10000 слов некоторого иностранного языка. В итоге пари он выиграл, но весь парадокс оказался в том, что он совершенно не мог общаться на том языке. Это еще раз напоминает известный факт, что язык, это не только слова. Таким образом, процессы устойчивого понимания речи являются узловыми на современном этапе речевого диалога.

4.2. Основные предпосылки

Группа исследования речи СПИИРАН в своей работе руководствуется следующими основными принципами:

- 1) Идеальным прототипом для моделирования понимания речи является система восприятия человека. Главными источниками знаний о свойствах этой системы являются: психоакустика речи [85], психология речевых ошибок и дефектов речи [86], области нейрофизиологии, посвященные обработке языка и речи [87]. Игнорируя эти свойства и факты, исследователи рискуют напрасно потерять время и средства.
- 2) Главная цель слушателя в процессе речевой коммуникации — это угадать намерение говорящего (но не вычислять истинность/ложность высказывания). Минимальной единицей речевой коммуникации является не фраза или какие-либо другие речевые выражения, а действие, которое выполняется с их помощью. Именно это действие и является смыслом высказывания [2].
- 3) Процесс восприятия речи содержит процедуры формирования гипотез звуков, слов и смысла, их квантитативное оценивание, соответствие конкретным априорным данным (акустико-фонетического, лексического, синтаксического, семантического и прагматического характера) и, наконец, выбор оптимального решения о смысловой интерпретации на основе некоторой интегральной оценки.
- 4) Попытка смысловой интерпретации должна быть всегда, несмотря на сомнения, возникающие на некоторых уровнях анализа.
- 5) Обработку информации синтаксического и семантического типов целесообразно проводить совместно, поскольку это соответствует природе человека [12]. Кроме того, согласно наиболее предпочтительной гипотезе, эта обработка в подсознании человека реализуется на базе механизма ассоциации [88], поэтому целесообразно использовать ассоциации и в моделях.
- 6) Поскольку общеизвестно, что для точного понимания речевого высказывания необходимо учитывать текущую ситуацию в прикладной области, модель понимания должна содержать ситуативную модель прикладной области.
- 7) Роль языковых правил в модели должна понижаться в процессе перехода от «правильной» речи к реальной спонтанной речи. Единственным правилом, которое выполняется во всех случаях без каких-либо исключений, это «так говорить правильно, поскольку так принято говорить в ситуации подобной данной». Оно действует в речи всегда и без всяких исключений.

4.3. Ассоциации вместо синтаксического и семантического разбора

Ассоциативный анализ фраз-гипотез в нашей модели основан на следующих предпосылках:

- 1) знания, которые принято называть синтаксическими и семантическими реализуются в подсознании человека с помощью механизма ассоциаций, причем оба вида знаний реализованы в едином процессе [89, 90];

- 2) межсловные ассоциации поддаются количественной оценке с помощью бинарных статистик [91,92] или на основе экспертных оценок [93, 94];
- 3) различные цепочки слов можно оценить по степени связности на основе межсловных ассоциаций [95].

Пусть $W = \{w_1, w_2, \dots, w_g, \dots, w_N\}$ — словарь.

Для каждой упорядоченной пары слов (w_g, w_h) зададим матрицу

$$A_{[N,N]} = \|a_{gh}\|.$$

размерностью $N \times N$, $a \geq 0$. Матрица создается статистически или экспертным путем с использованием дискретной, например, 4-бальной шкалы.

Для произвольной последовательности слов длиной L

$$f_n = w_{n_1}, w_{n_2}, \dots, w_{n_1}, \dots, w_{n_L}.$$

мы извлечем из A подмножество A^* , которое содержит коэффициенты для всех упорядоченных пар (w_g, w_h) :

$$A^* = \{a_{n_1, n_2}, a_{n_1, n_3}, \dots, a_{n_2, n_3}, a_{n_2, n_4}, \dots, a_{n_{L-1}, n_L}\},$$

$$|A^*| = C_L^2.$$

В результате ассоциативным показателем фразы f_n может быть нормированная по L сумма:

$$E_{\text{ass}}(n) = \frac{1}{C_L^2} \sum_{n_k = 1}^{L-1} \sum_{n_s = n_k + 1}^L a_{n_k, n_s}, \quad k < s.$$

4.4. Ситуативная модель. Можно ли понимать речь вне ситуативного контекста?

Известно, что полное понимание речи возможно только с учетом достаточно широкого ситуативного контекста, в соответствии с жизненным опытом собеседников и их профессиональными навыками. В нашей системе понимания, прагматический компонент реализован на основе «мягкого» количественного сравнения входных фраз-гипотез с образцами фраз, каждая из которых однозначно соответствует определенному действию в рамках текущей ситуации [96]. В результате этого сравнения каждая гипотеза получает количественную прагматическую оценку E_{pr} , которая существенна для интегральной оценки E (см. ниже), необходимой для семантической интерпретации входной фразы.

Система понимания включает операции распознавания, которые выполняются в условиях неопределенности, содержащейся во входном сигнале, а также в работе алгоритмических блоков высших уровней. Эта неопределенность может быть значительно уменьшена за счет учета закономерности прагматического характера.

Учитывая, что число ситуаций в предметной области является конечным, прагматическая обработка ведется на основе модели в виде диаграммы состояний, как ориентированного графа. Ее дуги являются переходами из ситуации в ситуацию, каждая дуга соответствует определенному смыслу и связана с подмножеством эквивалентных фраз, означающих определенное речевое намерение или конкретную команду.

Как можно видеть из рис. 11,

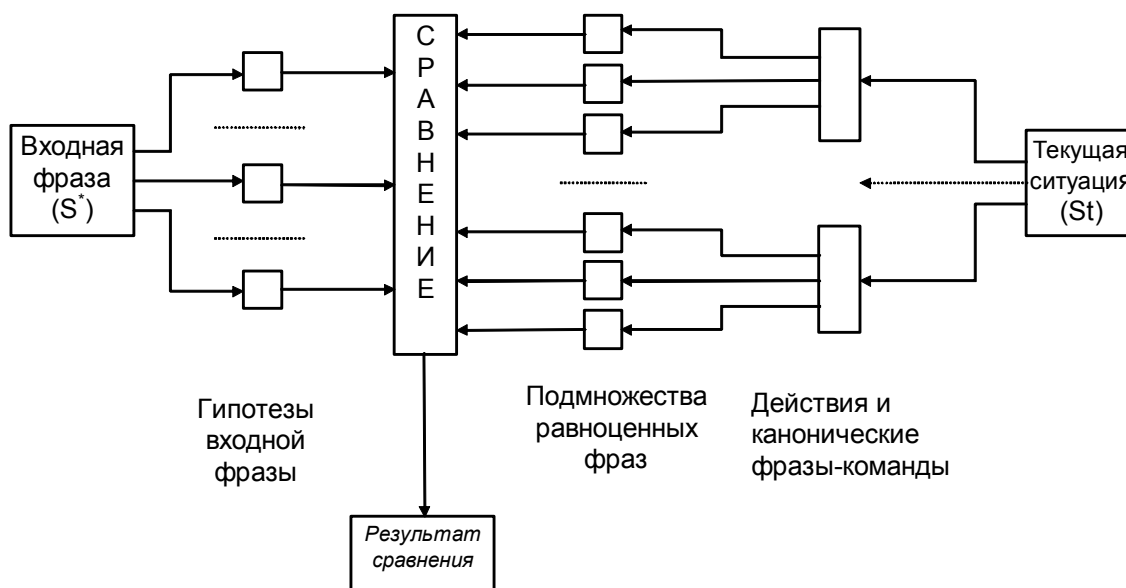


Рис. 11. Прагматическая оценка гипотез входной фразы

мы получаем оценки прагматического соответствия между входной гипотезой и конкретным актом путем количественного сравнения входной фразы со всеми вариантами команд, уместных в этой ситуации.

$$St = \{St_1, St_2, \dots, St_b, \dots, St_B\},$$

где B — это число ситуаций, в которые можно перейти из текущей ситуации.

Каждой ситуации St_b соответствует каноническая команда K_b :

$$St_b \rightarrow K_b.$$

Соответственно для каждого K_b существует определенное подмножество равноценных фраз

$$K_b \rightarrow \{K_{b1}, K_{b2}, \dots, K_{bj}, \dots, K_{bJ}\}.$$

Теперь только осталось определить смысловое расстояние между гипотезой входной фразы $f_n = H$ и канонической фразой $K_{bj} = K$. Оценка производится при следующих допущениях.

- Гипотеза H и команда K рассматриваются как подмножество слов, но не как строгие последовательности слов, поскольку в разговорной речи синтаксические нормы часто нарушаются.
- Каждое слово w_i из фразы команды имеет собственный смысловой вес v_i , который может быть заранее оценен экспертным способом.
 - Сумма всех весов словоформ является постоянной:

$$\sum_{i=1}^L v_i = \text{const.}$$

Теперь возможно представить фразу-команду K , как неупорядоченный набор пар <слово, его вес>:

$$K \rightarrow \{ \langle w_1, v_1 \rangle, \langle w_2, v_2 \rangle, \dots, \langle w_i, v_i \rangle, \dots, \langle w_L, v_L \rangle, \}, \quad L = |K|.$$

Практически невозможно оценить веса слов входного высказывания ввиду их непредсказуемости. Следовательно, гипотеза относительно входного высказывания должна быть представлена в виде:

$$H = \{w_1, w_2, \dots, w_i, \dots, w_M\}, \quad M = |H|.$$

Теперь рассмотрим семантическое различие подмножеств K и H . Резонно предположить, что оно зависит от конкретного лексического содержания K и H , а также от L , M , и V_i . Это расстояние количественно оценивается с использованием разностей множеств $A_1 = H \setminus K$ и $A_2 = K \setminus H$, и пересечения множеств $A_3 = K \cap H$ (рис.12).

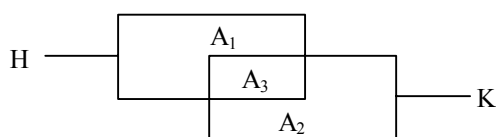


Рис. 12. Семантическое различие между гипотезой входной фразы и канонической фразой.

В результате подобных рассуждений была создана эмпирическим путем формула оценки прагматического различия между входной фразой и некоторым вариантом канонической фразы в виде:

$$D(H, K) = \frac{p_1 |A_1| + p_2 |A_2|}{L + M} \left(\sum_{i \in A_2} v_i + 1 \right).$$

Легко убедиться, что значение $D(H, K)$ возрастает с увеличением $|A_1|, |A_2|$ и суммой весов слов, входящих в подмножество A_2 . Для нормализации рассматриваемой функции по суммарной длине фраз применен множитель: $\frac{1}{L + M}$. Весовые коэффициенты p_1 и p_2 задаются экспертным путем. Все это согласуется со здравым смыслом.

Прагматическая оценка вычисляется для каждой гипотезы входной фразы. При этом гипотеза сравнивается с помощью вышеуказанной формулы со всеми вариантами канонических фраз, уместных в данной ситуации. Далее производится оптимизация, т. е. каждой гипотезе n ставится в соответствие минимальная из оценок, а также указатель b на соответствующее действие

$$E_{pr}(n, b, j) = \min_{n, b, j} D \{H_n, K_{bj}\}.$$

4.5. Интегральная обработка, что это?

На сегодняшний день нет достаточной ясности, что это такое, каким образом обрабатываются разные виды/уровни информации о речи и о прикладной

области, что может быть результатом такой обработки, в чем состоит ее эффект. Имеются только довольно обтекаемое (размытое) определение в работах Р. Шенка и М. Селфриджа [12, 22] которое опирается на гипотезу, что *синтаксический и семантический анализ (а также семантический и прагматический) выполняются в одно и то же время на основе одного и того же механизма*.

К сожалению, под это определение подпадают и такие модели, которые ничем не лучше обычных, последовательных моделей. Кроме того, требование одновременности представляется сомнительным т.к. оно исключает возможность интеграции, например, акустико-лексического уровня и семантико-синтаксического, поскольку эти уровни принципиально разновременны. Однако такая необходимость очевидна и уже успешно реализована в ряде моделей [40, 77].

В результате наших исследований проблемы интеграции, предложено такое определение: *интегральной обработкой является такая, которая позволяет взвешенно учитывать результаты частных видов обработки для возможности оптимального решения задачи понимания*.

Это определение приводит к структуре модели, показанной на рис. 13, которая выгодно отличается от привычной последовательной обработки (рис. 14).

Математическое представление интегральной обработки представлено ниже. Предположим, что каждый вид знаний 1,2, . . . n позволяет получить соответствующие оценки согласования входного высказывания с соответствующим видом знаний:

$$\left\{ \begin{array}{l} \text{Знание 1: } F_1(S) = 0; \\ \text{Знание 2: } F_2(S) = 0; \\ \dots \\ \text{Знание n: } F_n(S) = 0. \end{array} \right.$$

Например, на уровне распознавания слов это обычные DP- или СММ-оценки. Само собой очевидно, что нулевые оценки могут быть только в идеальном случае. Т.е., в общем случае при обработке реальной речи, правая часть должна содержать некоторые остатки:

$$\left\{ \begin{array}{l} F_1(S^*) = E_1; \\ F_2(S^*) = E_2; \\ \dots \\ F_n(S^*) = E_n. \end{array} \right.$$

Таким образом, вектор $E = E_1, E_2, \dots, E_n$ может характеризовать качество входной гипотезы. Его длина отражает степень отклонения входной гипотезы от некоторой гипотезы смысла и может служить базой для выработки оптимального решения по принципу «минимального отклонения».

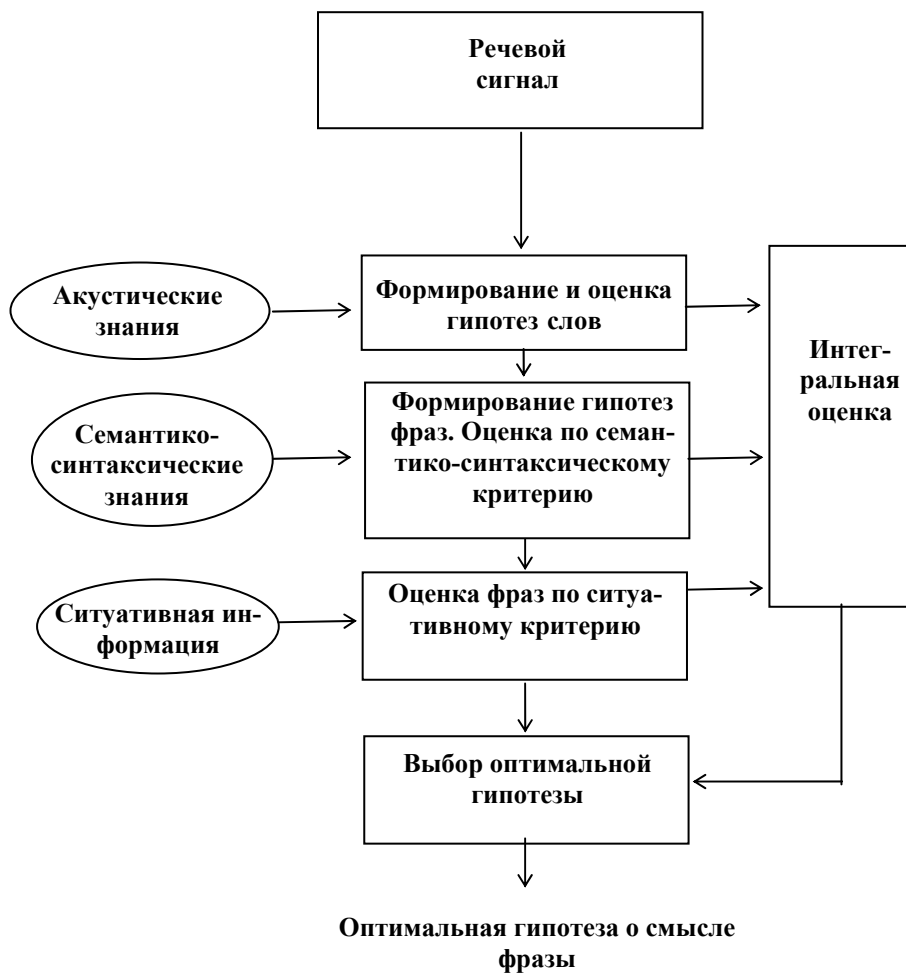


Рис. 13. Интегральная обработка речи



Рис. 14. Последовательный разбор речевого сигнала

4.6. Краткое описание экспериментальной модели понимания речи

В этой части мы кратко описываем работающую модель понимания речи разработанную для системы управления бытовым роботом. В этой модели мы используем квази-слитный ввод речи с использованием синтагм. Известно, что так называемая синтагма — это часть фразы — цепочка слов, объединенных некоторым критерием (дыханием, устойчивым смыслом и др.)

Мы используем синтагмы как группы слов тесно связанные процессом произношения. Например, *[Would_you_like] [to_go] [sightseeing?]*.

Базовая модель содержит следующие основные модули:

- 1) Модуль ввода и распознавания слов/синтагм. Он выполняет определение границ слов, их распознавание методом ДП, выдачу 10 лучших гипотез на следующий уровень обработки.
- 2) Модуль построения гипотез фраз, который использует данные предыдущего уровня и формирует всевозможные словосочетания (см. “example” ниже).
- 3) Модуль семантико-синтаксической обработки. Он использует ассоциативную базу данных, фрагмент которой показан ниже.

Номер и имя	1 - 10	11 - 20	21
1.go	4 4 4 4 4 1 3 2 4	3 4 4 3 1 4 4 1 1 1	1 1 1 1 1 1 1 1 1 1
2.to_the_living_room	3 1 4 1 1 2 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
3.living_room	3 1 4 1 1 2 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
4.please	4 4 4 4 4 4 4 4 4 4	4 4 4 4 4 4 4 4 4 4	4 4 4 4 4 4 4 4 4 4
5.to_the_bathroom	3 1 1 4 2 3 2 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
6.bathroom	3 1 1 4 2 3 2 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
7.go_to	1 3 4 4 3 4 3 2 4	4 2 4 4 3 4 4 2 4	3 1 1 1 1 1 1 1 1 1
8.the_bathroom	3 1 1 4 2 1 3 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
9.go_kitchen	2 1 1 4 1 1 3 1	2 2 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
10.kitchen	3 1 1 4 1 1 3 1	2 1 1 1 1 1 1 1 1 1	1 1 2 1 2 1 1 1 1 1
11.the_kitchen	3 1 1 4 1 1 3 1	2 2 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
12.to_the_bedroom	3 1 1 4 1 1 3 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
13.bedroom	3 1 1 4 1 1 3 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
14.the_bedroom	3 1 1 4 1 1 3 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
15.go_nursery	2 1 1 4 1 1 2 1	1 1 1 1 1 1 1 1 1 1	2 2 1 1 1 1 1 1 1 1
16.nursery	3 1 1 4 1 1 3 1	1 1 1 1 1 3 2 1	1 1 1 1 1 1 1 1 1 1

F1-Помощь F2-Сохранение F5-Печать F10-Выход Esc-Выход без сохранения

В этой базе данных каждая упорядоченная пара синтагм (слов) оценивается ассоциативной оценкой по 4-бальной шкале: 4 — отлично, 3 — хорошо, 2-удовлетворительно, 1- плохо. В соответствии с этими данными каждая входная фраза получает семантико-синтаксическую оценку согласно разд. 4.2.

- 4) Прагматический (ситуативный) модуль использует базу данных, фрагмент которой показан ниже.

Situation name	Name 1 of following situation	Phrase 1.1 Phrase 1.2	Weight of words Weight of words
	Name 2 of following situation	Phrase 2.1 Phrase 2.2

На основе этой базы данных по методу, описанному в разд. 4.3 вычисляется прагматическая оценка для каждой фразы-гипотезы.

- 5) Принятие решения осуществляется в соответствии с методом, описанным в разд. 4.4 с использованием выражения:

$$E_n = \left[\alpha_1 E_{ac}^2(n) + \alpha_2 E_{ass}^2(n) + \alpha_3 E_{pr}^2(n, j_n) \right]^{\frac{1}{2}},$$

где α_1 , α_2 , и α_3 являются весовыми коэффициентами и j_n означает соответствующее действие в рамках текущей ситуации. E_n показывает меру интегрального отклонения каждой гипотезы. Гипотеза с минимальной величиной E_n является результатом процесса интегрального понимания входного сигнала.

Для примера ниже приведены результаты распознавания/понимания одной из фраз приложения ROBOT.

EXAMPLE FOR INTEGRAL UNDERSTANDING

W o r d s r e c o g n i t i o n

go	174	the_kitchen	203
home	205	to_the_kitchen	209
to_boy	227	take	305
to_mother	244	the_bathroom	316
go_to	253	kitchen	318
bathroom	258	that	318
what	258	please	321
is_dog	267	to_the_bedroom	325
go_home	269	the_letter	326
to_father	272	a_letter	330

На уровне распознавания слов, как видно из этого листинга выводятся и имена 10 лучших слов (синтагм)–гипотез и их оценки, полученные методом динамического программирования (минимальная оценка — наилучшая гипотеза).

Phrase hypotheses estimation		Acc.	Ass.	Situat.
Int.				
go_to	the_kitchen	228	0	0(a)
228				
go	kitchen	246	0	0(b)
246				
go	to_the_bedroom	249	0	0(c)
249				

go_to 284	the_bathroom	284	0	0(d)
go_to 397	kitchen	285	0	112(e)
go 400	the_kitchen	188	100	112(f)
bathroom 401	please	289	0	112(g)
is_dog 406	please	294	0	112(h)
go 407	please	247	0	160(i)
to_mother 435	the_kitchen	223	100	112(j)

Здесь выведены 10 лучших входных фраз-гипотез со своими оценками: Acc — акустическая, Ass — ассоциативная, Situat — ситуативная, Int — интегральная.

S e n s e h y p o t h e s e s

- (a) go_to the_kitchen (1-4)
- (b) go kitchen (1-4)
- (c) go to_the_bedroom (1-5)
- (d) go_to the_bathroom (1-3)
- (e) go kitchen (1-4)
- (f) go_to the_kitchen (1-4)
- (g) go bathroom (1-3)
- (h) where is_dog (1-19)
- (i) go to_the_living_room please (1-2)
- (j) go_to the_kitchen (1-4)

Result – transfer from sit1 to sit4

Здесь приведены варианты решений о переходе из ситуации в ситуацию (1-4),(1-5) и т.п. Нужно отметить, что система поняла фразу совершенно правильно, несмотря на ошибку на уровне распознавания 1-го слова.

5. Заключение

Проведенный анализ исследований в области понимания речи и системы речевого диалога, а также опыт группы речевой информатики СПИИРАН позволяют высказать уверенность, что дальнейший прогресс можно ожидать за счет прорыва в следующих направлениях.

- 1) Создание более эффективных алгоритмов понимания речи на базе использования свойств живого интеллекта, т.е. таких феноменов, как ассоциативное восприятие, сквозное понимание, чувство ситуации, интегральная обработка и др.
- 2) Создание методов быстрой (по аналогии с человеком) адаптации к голосу, к акустической обстановке и конкретной прикладной задаче.

- 3) Создание компьютерной методологии быстрой подготовки пользователей в части приобретения навыков пользования микрофоном и устранения мелких дефектов речи.
- 4) Создание методов достаточно оперативного накопления статистической информации всех уровней, оценки разумного объема этой информации в зависимости от класса решаемых задач.

Литература

- [1] *Косарев Ю. А.* Естественная форма диалога с ЭВМ. — Л: «Машиностроение», 1989.
- [2] *Austin J. L.* How to do things with words. — New York :“Oxford University Press”, 1973.
- [3] *Chomsky N.* On certain formal properties of grammars. — Inform. Control 2, 1959. — pp. 137–167.
- [4] *Городецкий Б. Ю.* Компьютерная лингвистика: Моделирование языкового общения // Новое в зарубежной лингвистике. Вып. 24. — М.: Прогресс, 1989. — 5–32 с.
- [5] *Pfeifer R., Scheier Ch.* Understanding Intelligence. — Cambridge, MA: The MIT Press. Bradford Books, 1999.
- [6] *Налимов В. В.* Непрерывность против дискретности в языке и мышлении. — Тбилиси: Изд. Тбилисского университета, 1978.
- [7] *Косарев Ю. А.* Кодовая модель речевого сообщения. — Препринт ЛИИАН.Л., 1986. — 23 с.
- [8] *Fillmore Ch.* The Case for case. In Bach, Harms. Universals in Linguistic Theory. — New York, 1968.
- [9] *Попов Э. В.* Общение с ЭВМ на естественном языке. — М.: Наука, 1982. — 37 с.
- [10] *Schank R.* Conceptual Information Processing. — Amsterdam, North-Holland, 1975.
- [11] *Скороходько Э. Ф.* Семантические сети и автоматическая обработка текста. — Киев, 1983. — 112 с.
- [12] *Schank R., Birnbaum L., May J.* Integrating semantics and pragmatics. — «Quaderni di Semantica», 1985. — Vol. VI, no. 2.
- [13] *Lowerre B., Reddy D.* The Harpy speech understanding system. — Pittsburgh: Carnegie — Mellon University, 1976.
- [14] *Minsky M.* A Framework for Representing Knowledge. — Mass.: Artificial Intelligence Laboratory, Cambridge, 1974.
- [15] *Winston P.* Artificial intelligence. — London: Addison-Wesley Publishing Company, 1977.
- [16] *Schank R., Abelson R.* Scripts, Plans, Goals and Understanding. — Hillsdale, NJ: Lawrence Erlbaum, 1977.
- [17] *Винцюк Т. К.* Распознавание слов устной речи методами динамического программирования. — М.: Кибернетика, 1968. — №1, 15–22 с.
- [18] *Bellman R. E.* Dynamic programming. — Princeton, New Jersey: Princeton University Press, USA, 1957.
- [19] *Forsyth R.* Expert systems. Principles and case studies. — London: Chapman and Hall, 1984.
- [20] *Zadeh L.* «A fuzzy-algorithmic approach to the definition of complex or imprecise concepts». — In international Journal of Man-Machine Studies, 1976. — vol. 8, No. 3, pp.249–291.
- [21] *Shortliffe E.* Computer based Medical Consultations: MYCIN. — New York: American Elsevier, 1976.
- [22] *Selfridge M.* Integrated Processing Produces Robust Understanding. — “Computational Linguistics”, Vol. 12, №2, April-June 1986. — pp.89–106.
- [23] Представление и использование знаний / под ред. Х. Уэно, М. Исидзука. — М: Мир, 1989. —220 с.
- [24] *Zampolli A.* Interview for Elsnews on Language Resources Conference. — Elsnews № 7.3, 1998.
- [25] *Wang Ye-Yi.* A robust parser for spoken language understanding. — Budapest, Hungary: Eurospeech’99, 1999.

- [26] *Wang J., Wang H., Lee K., Huang C.* Domain-unconstrained language understanding Based on CKIP-Auto Tag, How-net, and ART // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [27] *Huang Y., Zheng F., Xu M., Yan P., Wu W.* Language Understanding Component for Chinese Dialogue System // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [28] *Lucke Helmut.* Interface of stochastic context-free grammar rules from example data using the theory of Bayesian belief // The Proc. of Eurospeech93, 1993. — pp. 1195–1198.
- [29] *Pearl J.* Probabilistic Reasoning in Intelligent Systems. — Morgan&Kaufmann, 1987.
- [30] *Minker W.* Stochastically-Based Natural Spoken Language Understanding Across Task and Languages. Spoken language Processing Group. — LIMSI-CNRS. EUROSPEECH'97, 1997 — p.p. 1423-1426.
- [31] *Dreyfus H. L.* What Computers Can't Do. A critique of artificial reason. — Harper & Row, Publishers, 1973.
- [32] *Swerts M., Litman D., Hirschberg J.* Corrections in Spoken Dialogue Systems // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [33] *Hirasawa J., Miyzaki N., Nakano M., Aikawa K.* New Feature Parameters for Detecting Misunderstanding in Spoken Dialogue System // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [34] *Rahim M., Pieaccini R., Eckert W., Levin E., Di Fabrizio G., Riccardi G., Kamm C., Narayanan S.* A Spoken Dialogue System for Conference / Workshop Services // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [35] *Wang H. M., Lin Y. C.* Coal-oriented Table-driven Design for Dialog manager // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [36] *Kurematsu A.,Akegam Y.,Burge S.,Jekat S.,Lause B., Maclaren V., Oppermann D., Schultz T.* VERBMOBIL Dialogues: Multifaced Analysis // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [37] *Zhang H., Xu B., Huang T.* How to Choose Training Set for Language Modelling // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [38] *Luo X., Franz M.* Semantic Tokenization of Verbalized Numbers in Language Modeling // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [39] *Miller G., Isard S.* Some Perceptual Consequences of Linguistic Rules, J. of Verbal Learning and Verbal Behavior, 1963. — 2, pp. 217–228.
- [40] *Jelinek F.* The Development of an Experimental of Discrete Dictation Recognizer // Proceedings of IEEE, No. 11, vol. 73, 1985
- [41] *Bahl L. R. et al.* Performance if the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task // Proceedings of ICASSP, 1995. — pp. 41–44.
- [42] *Leonardi F., Micca G., Militello S., Nigra M.* Preliminary results of multilingual interactive voice activated telephone service for people-on-the-move // Proceedings of EUROSPEECH'97, 1997. — vol. 4, pp. 1771–1774.
- [43] *Geutner P., Arevalo L., Breuninger J.* VODIS — Voice-Operated Driver Information Systems: A Usability Study on Advanced Speech Technologies for Car Environments // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [44] *Vilar J., Llorens D., Vidal E.* Experiments with Finite-State Models for Speech-Input Language Translation // Proceedings of SPECOM'96. — St-Petersburg, 1996. — pp. 59–63.
- [45] *Yokoo A., Sagisaka Y., Campbell N., Iida H., Yamamoto S.* ATR-MATRIX: Speech Translation System from Japanese to English // Proceedings of SPECOM'98. — St-Petersburg, 1998. — pp. 203–206.
- [46] *Pickles J. O.* An Introduction to the Physiology of Hearing. — New York: Academic press, USA, 1988.
- [47] *Markel J., Gray A.* Linear Prediction of Speech. — New York: Springer-Verlag, USA, 1980.
- [48] *Rabiner L. R., Schafer R. W.* Digital Processing of Speech Signals. — New Jersey: Prentice-Hall, Englewood Cliffs, USA, 1978.
- [49] *Davis S., Mermelstein P.* Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // Proceedings of ASSP'28, 1980. — pp. 357–366.
- [50] *Zwicker E., Terhardt E.* Analytical expressions for critical-band rate and critical bandwidth as a function of frequency // Journal of the Acoustical Society of America, 1980. — vol. 68, No. 5, pp. 1523–1525

- [51] *Rabiner L., Juang B.* Fundamentals of Speech Recognition. — New Jersey: Prentice-Hall, Englewood Cliffs, USA, 1993.
- [52] *Strom N.* Continuous Speech Recognition in the WAXHOLM Dialogue System. — STL QPSR, 1996. — pp. 67–96.
- [53] *Markel J., Gray A.* Linear Prediction of Speech. — Berlin, Springer-Verlag, 1976.
- [54] *Makhoul J., Raucos S., Gish H.* Vector Quantization In Speech Coding // Proceedings of IEEE, 1985. — vol. 73, No. 11, pp. 1551–1588.
- [55] *Johnson S. C.* Hierarchical clustering schemes. Psychometrika. — 1967. — 32, pp. 241–254.
- [56] *King B. F.* Step-wise clustering procedures // Journal of the American Statistical Association, 1967. — 62, pp. 86–101.
- [57] *MacQueen J. B.* Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. — 1967, pp. 281–297.
- [58] *Ball G., Hall D.* ISODATA, A Novel Method of Data Analysis and Pattern Classification. — (AD 699616) California, Stanford Research Institute, 1965.
- [59] *Sakoe H., Chiba S.* Dynamic programming optimization for spoken word recognition. — IEEE Trans. ASSP-34, 1986. — No. 1, pp. 52–59.
- [60] *Sakoe H., Chiba S.* Recognition of Continuously Spoken Words based on Time-Normalization by Dynamic Programming // J. Acoust. Soc. Japan, 1971 — 7, 9, pp. 483–490.
- [61] *Jelinek F.* A fast sequential decoding algorithm using stack. — IBM J. Res. Develop., 1969. — No13: 675–685.
- [62] *Levinson S. E.* "Structural Methods in Automatic Speech Recognition." — In Proceedings of the IEEE, 1985. — vol. 73, no. 11, pp. 1625–1650.
- [63] *Bakis R.* Continuous speech word recognition via centisecond acoustic states. — Washington: In Proc. ASA Meeting, 1976.
- [64] *Lee K. F., Hon H. W.* Large-vocabulary speaker-independent continuous speech recognition // Proc. IEEE Int Conf. On Acoustic, Speech, and Signal Processing, 1988. — pp. 123–126.
- [65] *Bahl L. R., Jelinek F., Mercer R.* A maximum likelihood approach to continuous speech recognition. — IEEE Trans. Pattern Anal. Machine Intell., 1983. — vol. PAMI-5, pp. 179–190.
- [66] *Viterbi A. J.* Error bounds for convolutional codes and an asymmetrically optimum decoding algorithm. — IEEE Transactions on Information Theory, 1967. — vol. IT-13, pp. 260–267.
- [67] *Sakoe H.* Two-Level DP Matching — A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. — IEEE Trans. ASSP-27, 1979. — No. 6, pp. 588–595.
- [68] *Myers C. S., Rabiner L. R.* A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition. — IEEE Trans. ASSP-29, 1981. — No. 2, pp. 284–297.
- [69] *Vintsiuk T. K.* Element-Wise Recognition of Continuous Speech Consisting of Words from a Specified Vocabulary. — Kibernetika, 1971. — No. 2, pp. 133–143.
- [70] *Bellegarda J., Silverman K.* Toward Unconstrained Command and Control: Data-Driven Semantic Interface // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [71] *Bonneau-Maynard H., Devillers L.* A Framework for Evaluating Contextual Understanding // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [72] *Akinori I., Chiori H., Masaharu K., Masaki K.* Language Modeling by Stochastic Dependency Grammar for Japanese Speech Recognition // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [73] *Seward A.* A Tree-Trellis N-best Decoder for Stochastic Context-Free Grammars // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [74] *Carpenter B., Lerner S., Pieraccin R.* Optimizing BNF Grammars through Source Transformations // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [75] *Potamianos A., Kuo H.* Statistical Recursive Finite State Machine Parsing for Speech Understanding // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [76] *Esteve Y., Bechet F., R. de Mori.* Dynamic Selection of Language Models in a Dialogue System // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [77] *Lin Y., Wan H.* Error-tolerant Language Understanding for Spoken Dialogue Systems // Proceedings of ICSLP'2000, Beijing, China, 2000.

- [78] *Wu C., Chen Y., Yang C.* Error Recovery and Sentence Verification Using Statistical Partial Pattern Tree for Conversational Speech // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [79] *Kuo H., Lee C.* Discriminative Training in Natural Language Call Routing // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [80] *Chou W., Zhou Q., Kuo H., Saad A., Attwater D., Durston P., Farrell M., Scahill F.* Natural Language Call Steering for Service Applications // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [81] *Lai Y., Lee K., Wu C.* Intention Extraction and Semantic Matching for Internet FAQ Retrieval Using Spoken Language Query // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [82] *Johnsen M., Holter T., Svendsen T., Harborg E.* Stochastic Modeling of Semantic Content for Use in a Spoken Dialogue System // Proceedings of ICSLP'2000, Beijing, China, 2000.
- [83] *Horiuchi Y., Arsushi F., Ichikawa A.* New WWW Browser for Visually Impaired People Using Interactive Voice Technology. — Budapest, Hungary: In Proc. Of Eurospeech'99, 1999. — pp. 2139–2142.
- [84] *Pokrovski N. B.* Calculation and Measurement of Speech Legibility. — Moscow: Svjaz., 1962.
- [85] Psychological and Psycho-physiological Research of Speech / Edited by Ushakova T. N. — Moscow: "Nauka", 1985.
- [86] *Lindsay P. H., Norman D. A.* Human Information Processing. — NY and London: Academic Press, 1972.
- [87] Cognition and the symbolic processes/Edited by Weimer W.. Palermo D. — Hillsdale, 1974.
- [88] *Oaksford M., Chater N.* Against logistics cognitive science. — In Mind & Language, 1991. — vol. 6, No. 1, pp. 2-37.
- [89] *Lyons J.* Introduction to theoretical linguistics. — Cambridge: At the University Press, 1972.
- [90] *Kravez L. G.* Quantitative Merkmale englischer Nominalverbindungen // Sprachstatistik. Mit zahlreichen Skizzen, Tabellen und Schemata im Text. Uebersetzt von einem Kollektiv unter Leitung von Lothar Hoffman. Wilhelm Fink, Muenchen/Salzburg, 1973. — pp. 252–264
- [91] *Danejko M. Maschkina L., Nechaj O., Sorkina W., Saharanda A.* Statiatische Untersuchung der lexikalischen Distribution der Wortformen // Sprachstatistik. Mit zahlreichen Skizzen, Tabellen und Schemata im Text. Uebersetzt von einem Kollektiv unter Leitung von Lothar Hoffman. Wilhelm Fink, Muenchen/Salzburg, 1973. — pp. 239–251.
- [92] *Deese J.* On the structure of associative meaning // Psychological review, 1962. — vol. 69, No. 2, pp. 161–175.
- [93] *Howes D.* On the relation between the probability of a word as an association and in general verbal usage // Journal of Abnormal and Social Psychology, 1957. — vol. 54, No. 1, pp. 75–86.
- [94] *Kosarev Yu. A., Jarov P. A.* Associations help to recognize words // Proceedings of DAGA-95, Saarbruecken, 1995. — pp. 979–982.
- [95] *Bekwith R., Fellbaum C., Gross D, Miller G.* WordNet: A lexical database organized on psycholinguistic principle // Using On-line Resources to Build a Lexicon. Zernic U. (ed.), NJ: Lawrence Erlbaum, Hillsdale, 1992.
- [96] *Kosarev Yu., Piotrowski R.* Synergetics and 'Insight' Strategy for Speech Processing. Literary and Linguistic Computing — Oxford University Press, 1997. — Vol. 12, pp. 113–118.
- [97] *Kosarev Yu.* Achievements and Challenges in Speech Dialogue with Computer. — St-Petersburg: Proc. Intern.Workshop "Speech and Computer", SPECOM'2000, 2000 — pp. 1–7.