V. L. Arlazarov, Речевой ввод/вывод как развитие человеко-машинных интерфейсов, *Informatsionnye Tekhnologii i Vychslitel'nye Sistemy*, 2004, Issue 2, 3–10

Речевой ввод/вывод как развитие человеко-машинных интерфейсов

В.Л. Арлазаров

Аннотация. Речевое общение человека с компьютером можно смело назвать технологией 21 века. Вопросам технической необходимости речевых интерфейсов и их научной и технологической готовности посвящена данная работа.

Введение

Развитие человеко-машинных интерфейсов в последнее время идет бурными темпами. Однако реализация одного из самых естественных способов взаимодействия человека и машины - общение посредством звучащей речи - до сих пор остается нерешенной задачей. Быстрое развитие компьютерных технологий и интенсивное расширение сферы использования персональных компьютеров не только создают необходимые технические предпосылки для развития и использования речевого интерфейса, но и делают эту задачу чрезвычайно актуальной.

Техническая необходимость

Техническая необходимость использования речевых каналов общения человека с компьютерными системами обусловлена следующими причинами:

- миниатюризация вычислительных систем уже сегодня сделала ввод и вывод самыми узкими местами в развитии мобильных устройств. Речевой канал является практически единственной перспективой нормального взаимодействия человека с компьютером;

- широкое распространение компьютерных систем в повседневной жизни диктует необходимость появления новых быстрых систем ввода/вывода информации, не требующих от неподготовленного пользователя специальных навыков;

- применение речевого общения с автоматизированными системами расширяет круг их потенциальных пользователей;

- добавление нового канала ввода/вывода улучшает мобильность, эргономику и "интеллектуальность" взаимодействия человека с компьютерными системами;

- в некоторых областях применения компьютерных технологий речевой канал взаимодействия является единственно возможным, например, речевое управление бортовыми устройствами, управление компьютером и другими техническими устройствами в экстремальных, опасных для человека, условиях;

- многократное увеличение объемов информации, циркулирующей по каналам связи, требует, в частности, и существенного уплотнения потоков речевой информации, передаваемой по цифровым каналам связи. Добиться существенного уплотнения можно с использованием речевых технологий. Например, имея модули распознавания и синтеза речи, можно создать вокодеры со скоростью до 100 бит/с, тогда как иные подходы даже в перспективе вряд ли позволят передавать речь со скоростью ниже 600 бод.

Перспективу использования речевого канала общения человека с компьютером делают доступной технический прогресс, высокие научные достижения в области исследования речи, а также успехи в программистской теории и практике. Ниже перечислены факторы, обеспечивающие техническую подготовленность к решению этой задачи:

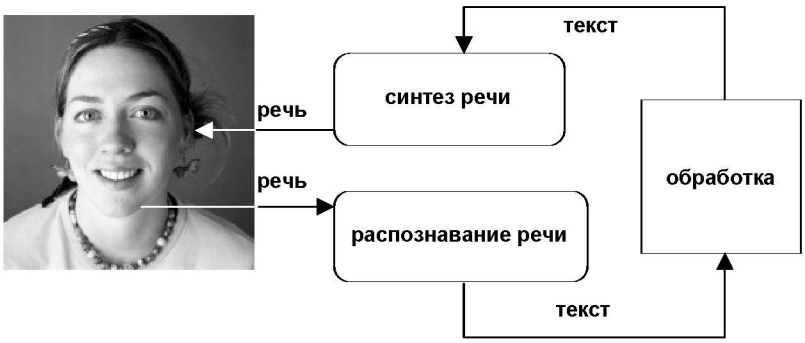

Рис. 1. Схема речевого общения человека с компьютером

- повышение производительности процессоров;
- возможность использования больших объемов оперативной памяти;
- появление надежных и компактных устройств цифрового ввода сигналов с высокой точностью и скоростью обработки на основе специализированных цифровых процессоров;
- повышение качества и доступность оконечной звуковой аппаратуры (микрофон, динамики, встроенные системы подавления шумов).

Перечислим факторы, связанные с научной основой рассматриваемой проблемы:

- наблюдаемый сегодня высокий уровень развития теоретической базы и практических исследований в мире;
- появление в 90-х годах и широкое распространение надежных методов распознавания, не требующих настройки на диктора и обеспечивающих работу в реальном времени;
- наличие широкого инструментария для исследований речи и разработки алгоритмов распознавания и синтеза речи, в том числе методы обработки и анализа сигналов, методы выделения параметров, крупные речевые корпуса;
- высокий уровень фонетических знаний о русской речи, что требует своего отражения в практических алгоритмах и системах.

Речевой канал ввода/вывода

В общем виде схема речевого общения человека с компьютером представлена на Рис. 1. Имеются преобразователь звучащей речи в сигнал, воспринимаемый компьютером, распознаватель, превращающий этот сигнал в текст, преобразователь, синтезатор речи и устройство озвучивания сигнала.

Если раскрыть подробнее блок обработки, то схема становится сложнее (Рис.2). Имеется узел управления диалогом, который специалисты по искусственному интеллекту обычно называют фреймом (не путать с фреймом в теории обработки сигнала - фрагментом анализа). Узел управления содержит словарь, используемый при распознавании и синтезе. Остальные блоки традиционны: база данных и переход в новый узел.

Наличие обратной связи между блоком обработки (узлом диалога) и системой распознавания речи позволяет существенно редуцировать задачу распознавания за счет уменьшения словаря и/или ожидания ограниченного числа семантических или синтаксических конструкций.

При разработке конкретных систем, разумеется, возникает множество различных задач. Те из них, которые специфичны для речевого общения, группируются вокруг двух основных проблем, решение которых в полном объеме дает ключ к построению огромного числа прикладных систем.

Первая проблема может быть обозначена как распознавание речи и характеристик голоса говорящего, вторая - синтез речи, включая восстановление похожести голоса. Сегодняшний уровень техники и научных исследований не позволяет решить эти задачи полностью. Поэтому они рассматриваются в определенных условиях, выявляющих специфику конкретных ситуаций (уровень шума, объем словаря, возможность на-

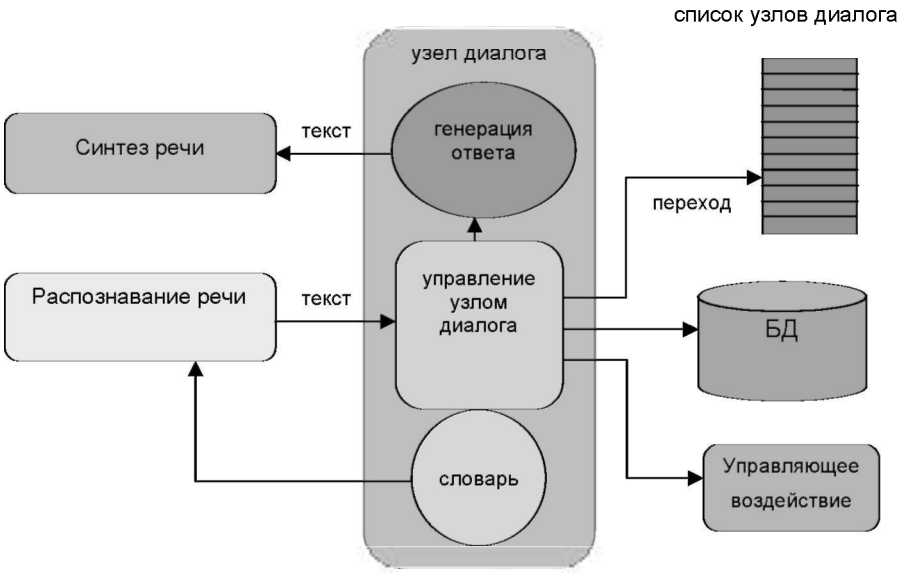

Рис. 2. Схема реализации речевого диалога человека с компьютером

стройки на диктора и т.п.). Однако из общности постановки следует, что вне зависимости от специализированных приемов, направленных на частные решения, имеется целый ряд методов, используемых практически всюду. Это касается первичной обработки и преобразований сигнала, методов дискретного и непрерывного распознавания, лингвистической обработки и др. Кроме того, при решении всех рассматриваемых проблем используется достаточно сложный инструментарий разработчика, большие акустические и фонетические базы данных и словари.

В каких системах может использоваться речевой интерфейс?

В качестве простейших примеров использования речевого интерфейса можно привести следующие:

- использование синтеза речи по тексту для озвучивания информации, получаемой из компьютера (возможно и без распознавания речи). Например - озвучивание ответов информационно-справочной службы мобильного оператора;
- распознавание небольшого количества команд. Например:
- распознавание цифр (голосовой набор номера телефона);
- мониторинг состояния машиниста поезда путем речевого подтверждения сигнала светофора (по этому примеру видно, что малый словарь, вообще говоря, не означает легкость решения задачи; так, в данной задаче требуется распознавание всего трех слов, но в условиях крайней зашумленности и с очень малой вероятностью ошибки);
- управление функциями бортового компьютера автомобиля: радио, климат-контроль, GPRS…(здесь словарь больше, а критичность ошибки разная - в зависимости от функции).

Более разветвленный речевой диалог возникает в таких задачах, как управление телефонным коммутатором, где присутствует дерево словарей и минимальная обратная связь, или заказ билетов по телефону, когда необходимость распознавать адрес может резко осложнить задачу.

В случае разветвленного речевого диалога мы имеем систему, управляемую заранее построенным графом диалога из управляющих шаблонов, при этом увеличивается размер словаря распознаваемых слов, однако в каждый момент словарь зависит от текущего состояния в графе диалога. Заметим также, что в интерактивных системах можно существенно повышать качество распознавания за счет повторного ввода сомнительных речевых сообщений и/или усложнения диалога.

Следующий уровень сложности речевого интерфейса представляют такие интеллектуальные речевые системы, как:

- система диктовки ("автоматическая пишущая машинка") - программа, которая распознаёт произвольное речевое сообщение и записывает его в текстовом виде;
- доступ к данным по телефону (Voice Portal);

- речевой ввод и озвучивание сообщений электронной почты;
- полнотекстовая индексация архивов с аудио- и видеоинформацией;
- интеллектуальные поисковые Интернет-системы с речевым общением на естественном языке;
- автоматический перевод звучащей речи на другой язык (здесь кроме распознавания речи на языке 1 и синтеза речи на языке 2 присутствует такой компонент, как автоматический перевод, по сложности даже превосходящий первые две задачи).

Следует также упомянуть задачи, решение которых связано с применением элементов речевого ввода в прикладных системах. Отметим при этом, что для их решения проблема распознавания речи дополняется задачей выделения параметров голоса говорящего:

- использование идентификации говорящего в качестве ключа доступа к системам с разграничением доступа;
- сжатие речи для хранения и передачи по каналам связи с целью создания низкоскоростных речевых преобразователей (вокодеров); с помощью распознавания речи с сохранением параметров голоса диктора речь преобразуется в текст с тем, чтобы после передачи по каналам связи синтезировать речь, обеспечив узнаваемость голоса;
- автоматические фонетические тренажёры (помощь в обучении иностранному языку);
- приспособления и компьютерные системы для помощи инвалидам (слепым, глухим, немым, парализованным).

Современное состояние речевых технологий в мире

Исследования в области автоматического синтеза слитной звучащей речи по тексту, введенному в ЭВМ, ведутся в мире много лет и показали большую научную и практическую сложность этой задачи. Сегодня существует несколько программ такого рода, которые показывают приемлемое качество озвучивания текста и могут найти практическое применение в специальных системах. Однако применение современных систем синтеза речи не становится повсеместным ввиду того, что они пока не обладают свойством достаточной естественности голоса и естественной просодической интонации "произнесения".

Обратная задача - распознавание текстовой информации по вводимой звучащей речи является еще более сложной задачей искусственного интеллекта. Однако сегодня представляется, что современная степень развития компьютерной техники и соответствующего периферийного оборудования, а также современные методы программирования, представления данных и распознавания образов дают все предпосылки для успешного решения этих задач.

В конце 90-х годов на мировом рынке программных продуктов появилось несколько программ, работающих в реальном времени и находящих некоторый коммерческий спрос в таких применениях, как передача информации в дискретной языковой форме ("диктовка"), ПК-интерфейсы, автоматическая телефонная служба, специальные промышленные и военные задачи. Однако технические характеристики этих продуктов не оправдали ожиданий пользователей, результатом чего стало банкротство целого ряда западных компаний, специализирующихся в распознавании речи. Эксперты прогнозируют увеличение рыночного спроса на подобные системы в ближайшем будущем, но подчеркивают, что это впрямую зависит от расширения их возможностей и улучшения точности распознавания. Среди главных требований, предъявляемых к таким системам, отметим в первую очередь увеличение объема словаря без существенной потери качества распознавания. Другими горячими точками здесь являются: распознавание в естественной акустической среде без направленного микрофона, естественный темп слитной речи (сейчас ошибка распознавания здесь примерно в 5 раз больше, чем для медленной или дискретной речи), независимость от голоса пользователя (ошибка распознавания примерно в 4 раза больше, чем в дикторо-зависимых системах), спонтанный диалог (ошибка распознавания примерно в 2 раза больше, чем при чтении

текста), проблемно-независимое распознавание (ошибка распознавания примерно в 2 раза больше, чем в проблемно-зависимом случае).

В настоящее время эта отрасль науки и технологии на западе снова активно развивается. Голосовые приложения подкрепляются новейшими стандартами, продвигаемыми консорциумом W3C и такими гигантами как Microsoft, Intel, IBM на основе языка XML. Существует ряд коммерческих программ распознавания и синтеза речи, в которых качество приближается к приемлемому. Увеличивается число систем, использующих эту технологию - заказ билетов, справочные службы, службы объявлений в аэропортах и т.д.

Среди западных компаний, предлагающих сегодня продукты распознавания речи, следует перечислить такие, как Scansoft (Dragon NaturallySpeeking, SpeechPerl, OpenSpeech), IBM (IBM ViaVoice), Philips (SpeechMagic Engine, Speech SDK), Microsoft (Speech API), Commodio (Qpoint command recognition, Qpoint dictate).

Ниже в таблице приведены характеристики некоторых коммерческих продуктов западных компаний.

Современное состояние речевых технологий в России

Рынок российских голосовых технологий до сих пор является предметом отдельных фрагментарных разработок, большинство которых не могут быть оценены как успешные. Единственная российская система диктовки "Горыныч", которая достигла рынка, была основана на системе фирмы Dragon и имела неудовлетворительное качество распознавания из-за недостаточно глубокого использования русской фонетики и лингвистики. Другие западные компании, которые пытались создать системы распознавания для русского языка, по разным причинам также не достигли успеха. В частности, бельгийская фирма Lernout&Hauspie в 2000 году начала

Название продукта	Тип распознавания	Языки	Скорость речи	Объем словаря	Заявленное качество	Цена
Dragon NaturallySpeaking	Диктовка + навигация	англ., нем., франц., японский	160 слов в минуту	250 000 слов + специализир. и пользов. словари	“Легко доводится до 99%”	$695+ $300 за каждый спец. словарь
IBM ViaVoice	Диктовка + навигация	англ., нем., итальян., японский		300 000 слов+ словари пользователей		$185
Dictaphone EXSpeech	Диктовка по шаблону для медицины	англ.			> 90% после настройки на диктора	$399
SpeechPerl	Навигация в телефонии	46 языков		Более 1 млн. слов		Зависит от кол-ва линий и объема словаря
OpenSpeech	Поддержка стандарта Voice XML			Более 1 млн. слов		
Philips Speech Processing Engine	диктовка	21 язык		128 000 слов – базовый словарь	95-98%	
Microsoft Speech API	навигация	англ.,япон. китайский				

разработку русскоязычного модуля распознавания, однако в настоящее время эта фирма обанкротилась. Прекратила финансирование аналогичного проекта из-за экономических трудностей и компания Intel.

Такая сложная задача, как разработка голосовых систем, требует финансовых ресурсов, которыми не обладает ни одна из российских фирм и научных организаций в отдельности. Более того, для решения задачи требуется объединение потенциала в нескольких областях науки: обработка сигнала, распознавание образов, фонетическая наука, компьютерная лингвистика.

При этом технологии преобразования текста в речь (text-to-speech) для русского языка гораздо ближе к готовности, многие организации демонстрируют системы с удовлетворительным качеством генерации речи уже сейчас.

Если же говорить о научных коллективах, занимающихся этой проблематикой в России, то в первую очередь следует назвать ИППИ РАН, ВЦ РАН, ИСА РАН, речевые группы филологического и механико-математического факультетов МГУ, Центр речевых технологий в Санкт-Петербурге, СПб Государственный Университет, СПИИРАН и др.

Научные подходы в распознавании речи

Распознавание речи как чисто научная проблема давно привлекло внимание ученых, и исследователи накопили множество методов, которые могут способствовать её решению.

При систематизации исследований по распознаванию речи можно выделить следующие основные подходы к построению систем распознавания речи:

1. подход, основанный на шаблонах, при котором входная речевая информация сравнивается с созданными ранее образцами и ищется наилучшее соответствие;
2. подход, основанный на использовании априорных экспертных знаний, при котором проводится попытка смоделировать способность человека распознавать речь по спектрограмме;
3. статистический подход, использующий вероятностные зависимости появления звуков и слов в речи;
4. непрерывный подход, при котором используются сети, состоящие из большого числа простых взаимосвязанных узлов, обученных распознаванию речи.

Простейшая система дикторозависимого распознавания изолированных слов может быть создана на основе сравнения с предварительно записанными шаблонами. Шаблоны, как правило, формируются из признаков, извлекаемых из речевого спектра. Для компенсации отклонений в скорости произношения применяется временная нормализация, производимая над последовательностью извлекаемых признаков. Нормализация может производиться при помощи алгоритма динамического программирования, позволяющего сгладить вариации в произнесении слова во временной области. Таким образом, данный подход основывается на непосредственном сравнении акустических параметров, соответствующих распознаваемому участку речи, с усредненными образцами, соответствующими распознаваемым словам. Образцы слов создаются в процессе обучения путем запоминания нескольких повторных произнесений слова и последующего усреднения акустических параметров. Основным достоинством подхода является его простота и надежность. Работа большинства систем, построенных таким образом, существенно зависит от диктора и условий записи. Применение данного подхода для распознавания больших словарей затрудняется сложностью в обучении системы, поскольку при обучении диктор должен несколько раз произнести все слова словаря. Кроме того, поскольку информация о каждом слове хранится отдельно, то при увеличении словаря потребности системы распознавания в памяти и вычислительных ресурсах могут быстро увеличиваться.

Подход, основанный на использовании экспертных фонетических знаний, предполагает содержательное исследование процесса произнесения или процесса восприятия и моделирование способности эксперта распознавать речь по спектрограммам и другим акустическим па-

раметрам. Системы, в которых делается попытка прямолинейной реализации такого подхода, обычно включают в себя стадии сегментации (т.е. выделение достаточно крупных фонетических элементов), построения гипотез о сегментах и последующий нечеткий вывод с учетом ограничений на длительность используемых фонетических элементов. В связи с большой изменчивостью речи, сложностью речевых моделей, существенным влиянием шумов и условий записи адекватное описание речи в виде экспертных правил затруднено. На практике в чистом виде такой подход не применяется. Тем не менее, в практических системах распознавания фонетические знания существенно используются при принятии решения об акустических параметрах, базовых единицах распознавания, при организации словарей и речевых баз данных, при подготовке материала для обучения систем. Использование фонетических знаний происходит опосредованно, при выборе структуры и алгоритмов системы распознавания.

Резкий скачок вперёд, произошедший в этом вопросе в последние годы, связан с широким применением так называемых статистических методов, в первую очередь - скрытых марковских моделей (HMM - Hidden Markov Model). Разумеется, важную роль здесь сыграло и увеличение скоростей ПЭВМ, позволившее их применять. Однако после недолгого периода эйфории наступило отрезвление. Пора быстрых успехов прошла, и дальнейшее продвижение практически застопорилось. Стало ясно, что одних только статистических методов решения задачи недостаточно.

Одним из возможных подходов является использование инженерии знаний "внутри" непрерывных методов. Реализация его является весьма трудной задачей, т.к. непрерывные методы как бы предполагают применение их на всем распознаваемом промежутке и любое вмешательство только портит дело. Между тем специальные методы кластеризации элементов могут позволить органично вписывать содержательные оценки в непрерывные схемы.

Другим примером гибридного подхода к моделированию и распознаванию речевых сигналов является схема, в которой оценки эмиссионных вероятностей делаются с помощью нейронных сетей, а моделирование и декодирование речевого сигнала выполняется традиционными методами. Результаты, полученные в этом направлении к сегодняшнему дню, дают уверенность в том, что на этом пути удастся построить более совершенные алгоритмы распознавания речи.

Заключение

Современное состояние проблемы речевых технологий дает основание полагать, что в самое ближайшее время в повседневную практику войдет использование речевого канала общения человека с компьютерными системами, в первую очередь, в области телефонии. Ожидается практическое внедрение таких технологий в автоматических информационно-справочных системах, обеспечивающих доступ по телефону, в телефонных системах заказа билетов и доставки товаров. Далее следует ожидать внедрения речевого канала в управление мобильными компьютерными устройствами и в перспективе - в управление бытовыми устройствами, создание полноценного голосового интерфейса с офисными приложениями и операционными системами.

Литература

1. Фланаган Д.Л. Речевое общение человека с машиной.- ТИИЭР, т.64, № 4, 1976.
2. Фланаган Д.Л. Анализ, синтез и восприятие речи. М., Связь, 1968.
3. Н.К. Обжелян, В.Н. Трунин-Донской. Речевое общение в системах "человек-ЭВМ". Кишинев, "Штиинца", 1985.
4. Lippmann R.P. Speech recognition by machines and humans. Speech Communication, 1997, v. 22, pp. 1-16.
5. Juang B.H., Chou W. Lee C.H. Statistical and discriminative methods for speech recognition. Automatic Speech and Speaker Recognition. Advanced Topics, Kluwer Academic Publications, 1996
6. Morgan N, Bourlard H. Neural Network for Statictical Recognition of Continuous Speech, Proceedings of IEEE, Vol 83, No 5, 1995, pp.742-770
7. Богданов Д.С., Брухтий А.В., Подрабинович А.Я., Усков А.В. Язык описания сценария диалога для речевого управления / В сб. "Развитие безбумажной техноло-

гии в организационных системах". М., Эдиториал УРСС, 1999.
8. Кривнова О.Ф., Зиновьева Н.В., Захаров Л.М. Программный синтез русской речи (синтезатор "АГАФОН") // Труды Международного семинара по компьютерной лингвистике и ее приложениям. Казань, 1995.С. 121-128.

Арлазаров Владимир Львович. Родился в 1939 году. Окончил механико-математический факультет МГУ им. М.В. Ломоносова. Доктор технических наук с 1987 года. Специалист в области математического обеспечения ЭВМ и искусственного интеллекта. Автор более 60 научных работ, в том числе монографий. Заведующий отделом Института системных исследований РАН, профессор Московского физико-технического института.