СИСТЕМА АКТИВНОГО ДИАЛОГА «ЧЕЛОВЕК-КОМПЬЮТЕР» С РУССКОЯЗЫЧНЫМ ГОЛОСОВЫМ ИНТЕРФЕЙСОМ

Использование речи при общении человека с различными справочно-информационными системами и управлении машинами и различными системами становится более и более насущной необходимостью. По оценкам экспертов американских исследовательских компаний Cahners In-Stat и Datamonitor объем рынка речевых технологий к 2005-2006 году возрастет до 3-5 млрд. долларов. В последнее время появилось достаточно большое количество приложений, использующих голосовой интерфейс. Большинство современных разработок в области распознавания и синтеза речи нацелены на решение следующих задач:

- Общение со справочно-информационными системами.
- Управление машинами и различными системами.
- Управление бытовой аппаратурой

По характеристикам голосовых интерфейсов они подразделяются на две большие группы:

1. Голосовой интерфейс работает без предварительной настройки, но распознается небольшое число слов (как правило, меньше 50, обычно 10 - 20).

2. Голосовой интерфейс распознает большое количество слов (до 30 000), но для получения приемлемого качества распознавания (90-92%) требуется длительная настройка на голос пользователя. При этом, как правило, необходима частая корректировка, особенно если по той или иной причине параметры голоса меняются (простуда, хрипота, много говорил, употребление алкоголя, курение).

Наша разработка основана на применении следующих, разработанных нами ноу-хау технологий:

- нового способа получения первичного описания речи с использованием новой математической модели звукового тракта человеческого уха;
- новой оригинальной модели нейронной сети для распознавания слов.

По сравнению с нижеописанными зарубежными программами можно отметить следующие преимущества нашей разработки:

1. **Использование новой технологии распознавания речи, которая позволяет распознать речь любого диктора без предварительной адаптации с точностью 95-98% на словаре 5-10 тысяч слов. Это крайне важно для различных информационно-справочных и других систем, работающих с большим количеством пользователей.**
2. **Эта технология может быть использована для обучения речи глухих людей, людей с дефектами произношения и при изучении иностранных языков, благодаря использованию нового способа получения первичного описания речи .**
3. **Позволяет легко интегрировать модули распознавания/синтеза речи в любые программные приложения.**
4. **Возможность использования различных справочно-информационных систем с помощью обыкновенного телефона.**

Система активного диалога «человек-компьютер» с мультимодальным (голос, изображение, текст) интерфейсом будет включать в себя:

1. Подсистему распознавания/синтеза естественной речи для голосового общения пользователя с информационно-поисковой системой[1]. **Распознавание речи ведется без предварительной настройки на диктора, охватывая широкий диапазон голосов — от мужского баса до детских.** Для этого на фирме разработаны:
   - способ инвариантного описания фонем
   - оригинальная нейронная сеть, классифицирующая определенные сочетания фонем как слова из заданного словаря[2].

2. Контекстно-зависимую диалоговую подсистему общения пользователя с интеллектуальным оборудованием. Такая система путем ведения диалога с пользователем старается максимально удовлетворить запросы пользователя. Диалог ведется в дружественной манере на естественном языке. Учитывая, что создать универсальное средство ведения диалога — задача пока непосильная, диалоговая подсистема настраивается на определенную прикладную задачу (электронная коммерция, справочно-информационная система, резервирование и продажа билетов, новостная служба и т.д.). Разработан оригинальный алгоритм диалоговой системы, обеспечивающей ведение предметно — настраиваемого, активного гибкого диалога[3].
3. Подсистему управления мультимодальным интерфейсом, позволяющим вести комбинированный диалог (выдавать/получать информацию в виде голоса, текста, изображения). Средства подсистемы управления должны обеспечивать преобразование информации из одного вида в другой. При ведении диалога пользователь может применять в качестве терминала офисный/домашний компьютер, подключенный к сети Интернет, специальный терминал, подключенный к серверу информационно-поисковой системы, мобильный или обычный телефон. При этом есть возможность сделать запрос с одного терминала (например, с мобильного телефона), а необходимую информацию получить на другой (например, как почту для компьютера).

Диалоговая система может быть использована как часть программного комплекса, предназначенного для резервирования и продажи билетов, электронной коммерции, а также послужить основой справочно-информационной системы, как для обслуживания клиентов, так и для внутри корпоративного использования.

Речевой интерфейс «человек-компьютер» обладает рядом бесспорных преимуществ:

- оперативностью и естественностью общения с интеллектуальными информационными системами;
- минимумом специальной подготовки пользователей;
- возможностью управления объектами когда «руки заняты» или пользователь является инвалидом;
- возможностью общения и управления интеллектуальными системами, приборами и устройствами по телефону.

Необходимо отметить, что в процессе исследований разработан ряд демонстрационных программ:

- Голосовой калькулятор — для демонстрации дикторонезависимого распознавания слов русского языка из ограниченного словаря;
- Диалоговая система для покупки авиабилетов (текстовый вариант);
- Голосовой интерфейс для Интернет-броузера;
- Комплекс игровых программ для обучения правильному произношению, превосходящий по качеству обучения аналогичную по назначению разработку фирмы IBM — SpeechViewer. При незначительной доработке графики интерфейса этот комплекс сам по себе может быть рыночным продуктом (стоимость пакета SpeechViewer около 1000$).

В ходе реализации проекта предполагается получить следующие продукты:

- Пакет программных инструментов(SDK) для встраивания голосового интерфейса и диалоговой подсистемы в разрабатываемые пользователем программные продукты (возможно на основе VoiceXML), включающий:
  - Подсистему распознавания/синтеза естественной речи для голосового общения пользователя с информационно-поисковой системой.

- Контекстно-зависимую диалоговую подсистему общения пользователя с интеллектуальным оборудованием.
- Подсистему управления мультимодальным интерфейсом, позволяющим вести комбинированный диалог (выдавать/получать информацию в виде голоса, текста, изображения).

А также для общего пользования:

- Комплекс игровых программ для обучения правильному произношению.
- Программу преобразования звуковых сигналов (в т.ч. речи) на основе инвариантного первичного описания звука для так называемых «Koehler implant» — протезов слухового тракта.

В последнее время появилось достаточно большое количество приложений, использующих голосовой интерфейс. Наиболее часто голосовой интерфейс используется для систем диктовки текстов. Встречаются телефонные сервисы, так называемые CALL-центры и новостные системы, использующие распознавание слов (команд) из ограниченного словаря и диалог по жестко фиксированному дереву вопросов-ответов. Несколько фирм, активно занимающихся проблемой распознавания речи, выпускают программный инструментарий для разработки таких систем на основе языка голосовой разметки VoiceXML.

Однако не известна какая-либо программная система, в которой были бы одновременно реализованы приведенные выше три составных части:

- Подсистема распознавания/синтеза естественной речи
- Подсистема ведения активного диалога
- Подсистема управления многомодальным интерфейсом

Основные идеи

В ходе реализации проекта особое внимание уделяется разработке алгоритмов и методов анализа речевых сигналов, направленных на получение инвариантного к голосам различных дикторов и акустическому окружению параметрического описания речевых сигналов. Известные способы предварительной обработки речевых сигналов используют в качестве основы методы спектральной обработки сигналов: исходный сигнал дискретизируется, обрабатывается с помощью преобразования Фурье, затем определяются параметры нескольких первых гармоник, несущих основную информацию об огибающей спектра (спектральное описание), а далее, с помощью набора эталонов произношения и скрытой марковской модели(Hidden Markov Model-HMM), определяется наиболее вероятно произнесенное слово.

Предлагаемые для реализации проекта алгоритмы и методы основаны на максимальном использовании принципов восприятия речи слуховым трактом человека. Это обусловлено рядом объективных факторов.

Если общую задачу распознавания речи представить как две независимые последовательно решаемые части, то первая часть — это получение первичного описания речевого сигнала, а вторая часть — это анализ изменений первичного описания во времени. При восприятии речи человеком первая задача решается слуховым трактом, а вторая — нейронной сетью мозга.

Для успешного решения всей задачи распознавания необходимо выработать объективные критерии правильности принятия тех или иных решений. В первой части задачи в качестве критерия используется надежность распознавания стационарных участков гласных фонем, произнесенных различными дикторами (мужские, женские и детские голоса), а при решении второй части — надежность распознавания отдельных слов, произнесенных разными дикторами, при этом в том и другом случае используется только одно эталонное описание (один кластер).

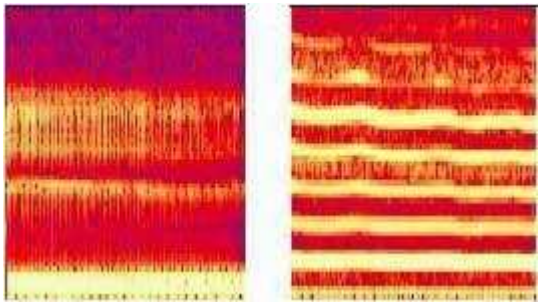
При получении первичного описания в качестве базиса при разработке программного обеспечения используется не набор идеальных математических функций (как, например, используется разложение в ряд Фурье при использовании спектрального подхода для получения первичного описания), а модель механического осциллятора. При анализе медленно меняющихся, строго периодических сигналов разница между предложенным методом и известными —

минимальна. Значительные отличия проявляются при анализе сложных сигналов, содержащих несколько резонансных максимумов (формант) и при анализе переходных процессов.

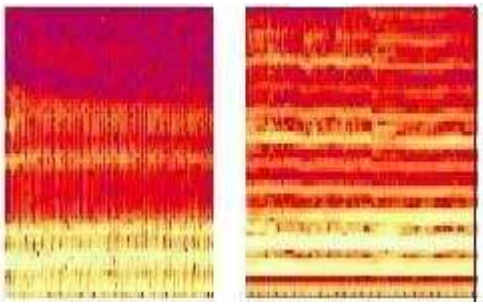
Исследователями уже давно высказывалось предположение о зависимости положения формант в спектре речевого сигнала от частоты основного тона. Сотрудникам фирмы «Суперкомпьютерные системы» удалось определить эту зависимость и использовать ее при распознавании.

В исследованиях по распознаванию речи установлен следующий факт: если подвергнуть речевой сигнал клиппированию (ограничению амплитуды сигнала постоянным значением), то, несмотря на столь значительное искажение, человек практически его не ощущает. На фирме смоделировали этот эффект и также используют эту модель при анализе.

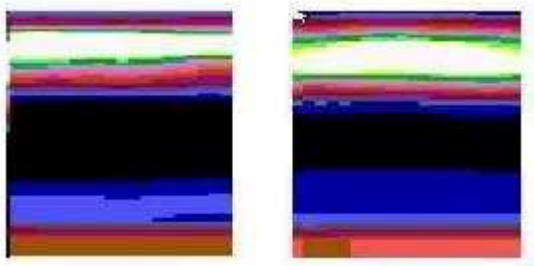
Известные методы позволяют добиться хороших результатов при распознавании речи от одного диктора, но при распознавании речи от нескольких дикторов надежность резко падает. Большая разница в надежности распознавания речи от одного и от нескольких дикторов, по-видимому, объясняется тем, что известные варианты первичного описания речевого сигнала значительно отличаются от того описания, которое является результатом обработки речевых сигналов в ушном тракте человека. Полученное на фирме инвариантное описание практически не зависит от индивидуальных особенностей голоса диктора и одинаково описывает как мужской бас, так и детский голос (внизу слева и справа на рис.1).



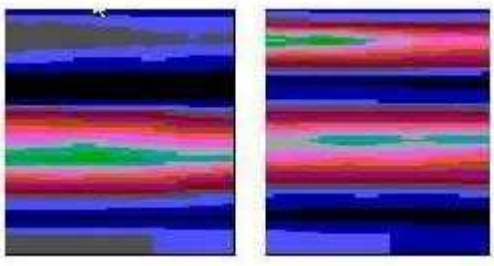
Стандартное спектральное описание русского звука **«И»** (английское **«E»**), которое было получено от двух различных дикторов.



Стандартное спектральное описание русского звука **«А»** (английское **«R»**), которое было получено от двух различных дикторов.



Здесь показаны те же звуки, что и на картинке вверху слева, но полученные при помощи нашей технологии. Легко заметить, что они очень похожи и практически незаметно, что они получены от различных дикторов.



Здесь показаны те же звуки, что и на картинке вверху справа, но полученные при помощи нашей технологии. Легко заметить, что они очень похожи и практически незаметно, что они получены от различных дикторов.

рис.1

В большинстве современных систем распознавания речи используется процедура адаптации эталонов распознаваемых элементов речи к голосу и манере произнесения конкретного пользователя, а также акустическим условиям окружающей среды, типу используемого микрофона и пр. Таким образом, для того, чтобы настроить систему пользователь должен прочитать специальные тексты, предлагаемые компьютером. При этом обучение желательно

проводить в тех же внешних условиях, в которых данная система будет затем эксплуатироваться. Это предполагает, как минимум, наличие высококачественного микрофона, расположенного на фиксированном расстоянии ото рта диктора и отсутствие окружающего шума. После этого надежность распознавания речи пользователя, на голос которого настроена система, не превышает 92-95%. При нарушении этих условий (например, при распознавании речи из телефонной линии) надежность распознавания уменьшается до 80% и ниже. Но даже при соблюдении всех перечисленных условий невозможно сохранить высокую надежность распознавания с течением времени, так как тембр голоса каждого диктора постоянно и существенно изменяется. Такие изменения зависят от того, как долго говорил диктор, его эмоционального состояния, ел ли диктор мороженое или пил горячий кофе и множества других субъективных факторов. ***Используемое в проекте инвариантное описание речевого сигнала моделирует процесс первичной обработки речи слуховым трактом человека и именно поэтому позволяет добиться максимальной надежности распознавания без адаптации к голосам конкретных пользователей системы и внешним факторам.***

При реализации второй части общей задачи распознавания речи — анализе речевого сигнала во времени необходимо суммировать информацию об отдельных звуках речи, последовательность которых и составляет речевое высказывание (синтагму). С этой целью наиболее широко используется метод HMM (Hidden Markov Modelling), который в настоящее время позволяет получить максимальную надежность распознавания. Суть этого метода состоит в том, что речь рассматривается как случайный процесс, т.е. внутренние процессы обработки речевого сигнала скрыты (hidden) от исследователя. Исследователь анализирует только внешние проявления этого процесса и, накапливая статистический материал, фактически «угадывает» распознаваемую речь. Надо признать, что до определенного предела эти методы достаточно эффективны, однако они не могут достичь надежности распознавания речи, сравнимой с надежностью распознавания речи человеком. Предел возможностей этих методов ограничивается тем, что нельзя априори учесть все возможные варианты изменения речевого сигнала.

Разработанная нейронная сеть моделирует процессы обработки речевой информации человеком, при этом сеть моделирует различные нелинейности и переходные процессы. Принцип расчета принадлежности речевого образа к тому или иному классу на нашей нейронной сети позволяет учесть все возможные варианты изменения речевого сигнала, к которым инвариантно человеческое ухо. Причем эта инвариантность обеспечивается даже в том случае, когда в качестве эталона используется только одно описание последовательности звуков в отличие от эталонов конкурентных аналогов, при формировании которых требуется задание множества различных вариантов произнесения каждого слова. Найденные методы позволяют обеспечить распознавание на основе только стандартной фонетической транскрипции, что принципиально упрощает задание эталонов для любого языка, используя лишь транскрипцию слов.

При распознавании синтагм применяемые в проекте методы позволяют полностью использовать просодическую информацию (интонацию и ударения) а также фонетические правила языка.

При распознавании речи в таких языках как русский практически невозможно задать априори все возможные варианты произнесения из-за большого количества предлогов, приставок, суффиксов и окончаний (словоформ). ***Создаваемая специализированная база данных с речевой информацией и специальный граф распознавания позволяют реально решить эту задачу.***

Очень важная составляющая проекта — система ведения диалога.

Для его ведения при реализации проекта использовано представление предметной области в виде многопараметрического пространства. Суть диалога при этом заключается в достижении точки (или некоторого подмножества точек пространства) с подходящими для пользователя параметрами, путем последовательного уточнения этих параметров. Далее значение найденной точки (точек) предъявляется пользователю для проведения с ними соответствующих операций. Так, например, для покупки авиабилета пользователь должен сообщить системе пункт назначения, желаемые дату и время вылета (прилета), тип места, авиакомпанию и т.д. Эти параметры в ходе диалога определяются, уточняются, возможно изменяются. Пользователь может определить

некоторые параметры как приоритетные. Тогда, в случае, если пожелания пользователя нельзя удовлетворить, система предложит альтернативное решение с максимально приближенными к требованиям пользователя параметрами, причем приоритетные параметры будут изменены в последнюю очередь..

В ходе диалога производится частичный семантический анализ высказываний пользователя с учетом предметной области диалога. Разбор предложений (высказываний) производится с использованием так называемых шаблонов, в состав которых входят:

- шаблоны параметров (основных глаголов и основных существительных, определяющих предметную область диалога);
- шаблоны определяющие реакцию пользователя в процессе диалога — согласие, не желание и вопрос;
- шаблоны предлогов, частиц, союзов и небольшого числа общеупотребительных словосочетаний, что позволяет определить отношение пользователя к вариантам, предложенным диалоговой системой.

По сути дела, все вместе — это набор исключительно прагматических образований, позволяющие решать предметно-ограниченную задачу, хотя часть шаблонов не зависит от предметной области, и следовательно, они могут служить общими шаблонами для разработок диалоговых систем для другой предметной области. Для дальнейшего разбора предложения не используется синтаксический и лексический анализ, так как ведется свободный, вербальный диалог, который славится своей грамматической корявостью и несоблюдением синтаксических правил. Для определения смысла, заключенного в предложении, разработан алгоритм, который сначала выделяет «слева-направо» основные существительные и затем, в соответствии с вышеуказанными шаблонами, определяет для каждого из них логически связанные с ними глаголы. Исходя из полученной информации, формируются параметры поиска. Основной акцент при разработке алгоритма разбора ставится на способность работать с большим числом последовательно идущих высказываний, в которых пользователь определяет параметры обсуждаемого предмета. То есть, по сути — это разбор сложносочиненного предложения. При этом используются в основном несколько эмпирических правил. Например, если набор из основных существительных (которые могут быть «разбавлены» другими словами) окружен с двух сторон глаголами не желания, то эти существительные вычеркиваются из поискового списка. Или, если набор основных существительных находится между шаблоном не желания, и шаблоном вопроса, за которым не следует основных существительных, а некоторые из этих основных существительных обсуждались ранее, то не упоминаемые ранее основные существительные из набора будут параметрами для дальнейшего поиска (таким образом, используется информация, накопленная в ходе диалога). Эмпирические правила также естественным образом базируются на объединительных и разделительных свойствах союзов.

Мультимодальное ведение диалога предполагает предоставить пользователю возможность вести диалог в желаемом варианте: например, пользователь говорит в микрофон, а компьютер выдает ответы на дисплей; пользователь ведет диалог с компьютером по телефону, а конечная информация посылается ему в виде e-mail'a, и т.д. Причем предусматривается возможность использования так называемого профиля пользователя — постоянно накапливающейся информации о предпочтениях и особенностях работы: один предпочитает голосовое управление просмотром биржевых сводок, второй — постоянно посещает сайты с анекдотами и любит, когда их ему читают, вечером пользователь прослушивает новости, а утром-сводку погоды и т.д., т.е. часть информации по ведению диалога может быть получена по инициативе системы, сразу же после идентификации пользователя и без его участия. Естественно, пользователь в любой момент может переключиться на другой способ общения.

В ходе проекта так же предполагается разработать способы преобразования информации в ходе диалога в зависимости от того, в каком виде она находится в базе данных, и от пожеланий пользователя.

Области применения

Диалоговая система предназначена для обмена информацией между человеком и различными программными системами, включающими ее как свою подпрограмму.

Она может использоваться для создания:

- голосовых Internet-порталов (телефонных сервисов),
- систем автоматического резервирования и продажи билетов,
- электронной коммерции,
- справочно-информационной системы, предназначенной как для обслуживания клиентов, так и для внутрикорпоративного использования.

Кроме того, голосовой интерфейс может быть встроен в различные приборы и устройства, для того чтобы:

- облегчить управление ими;
- дать возможность управлять ими инвалидам;
- дать возможность прибору голосом выдавать текущую информацию или предупреждения.

Еще один способ использования диалоговых систем — обучение правильному произношению. Это очень актуально для:

- глухих и слабослышащих;
- детей с дефектами речи;
- изучения иностранных языков;
- людей, для которых трудовая деятельность непосредственно связана с владением голосом (актеры, дикторы, преподаватели).

Примечания

1. Надо отметить, что предлагаемый подход инвариантен к языку. Поэтому при определенных доработках программные инструменты, полученные в ходе реализации проекта, могут быть использованы для любого из языков (по крайней мере, романской группы).
2. несколько демонстрационных программ, подтверждают правильность выбранного технического решения.
3. Создано несколько демонстрационных программ.