1. Основные задачи и ключевые понятия речевого интерфейса

К л ю ч е в ы е п о н я т и я :
речевой интерфейс, синтез речи, распознавание речи, понимание речи, распознавание голоса, компьютерное клонирование голоса и дикции, форманта, вокодер, речевой интерфейс, речевая система, система автоматического синтеза речи, синтезатор речи, система автоматического распознавания речи, мультимодальная система, мультимедийная система, диалоговая система, интеллектуальная система.

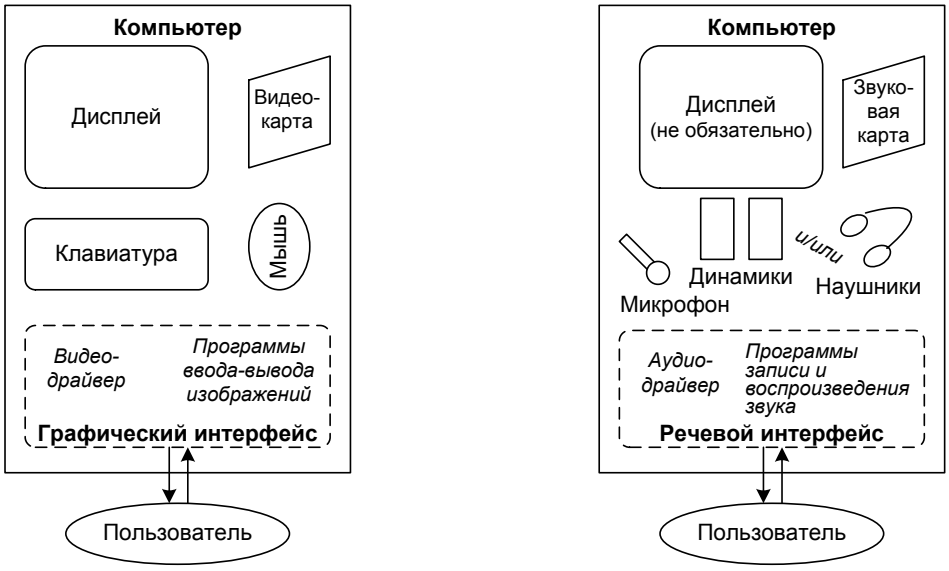
Повсеместное внедрение средств вычислительной техники в различные сферы человеческой деятельности делает разработку речевого интерфейса для взаимодействия с ЭВМ очень актуальной. Речевой интерфейс обладает рядом бесспорных и очевидных преимуществ:

- оперативность и естественность;
- минимум специальной подготовки пользователя;
- возможность управления объектом в темноте, за пределами его визуальной видимости (в частности, с использованием существующей телефонной сети);
- возможность использования одновременно ручного (с помощью клавиатуры) и речевого ввода информации;
- обеспечение мобильности оператора при управлении объектом и др.

Сказанное не означает, однако, что речевой способ человеко-машинного общения целиком заменит традиционные способы ввода-вывода информации [2] (***Жданович В.М..1991кн-Технич_С_ЭВМ***). Наряду с другими средствами он активно способствует дальнейшей интеллектуализации человеко-машинных систем. Результаты, полученные в области разработки речевого интерфейса, становятся доступны широким слоям населения. В настоящее время нам уже известны примеры широкого использования речевого интерфейса, например, в области телефонии, бытовой техники и др. Подробнее об этом будет сказано ниже.

Несмотря на большую актуальность, далеко не все задачи разработки речевого интерфейса в настоящее время можно считать решенными. Проблема разработки речевого интерфейса является достаточно сложной и комплексной, что требует от разработчика знаний в различных предметных областях (см. введение).

Речевой интерфейс является **аппаратно-программным комплексом**, так как для его реализации требуется использование внешних (дополнительных) по отношению к компьютерной системе аппаратных средств (микрофон и средства вывода звука, например, наушники, а также звуковая плата). Это обстоятельство накладывает дополнительные требования не только на уровне разработки компьютерных программ, обеспечивающих компьютер способностью говорить и слышать, но и на аппаратном уровне, от которого зависит, в частности, качество воспроизведения звука, оперативность обработки информации и др.



Р и с у н о к 1 . 1 . Отличие речевого интерфейса от графического («традиционного»)

Как показано на рис.1.1, речевой интерфейс имеет ряд отличий от ставшего уже традиционным графического интерфейса. В связи с этим меняется как технология разработки самого интерфейса, так и технология взаимодействия пользователя с компьютером. Это следует учитывать при разработке интеллектуальной системы, поскольку одними из важнейших показателей качества интерфейса компьютерной системы [3] (***Торрес Роберт Дж.2002кн-Практ_Р_по_П***) являются его **удобство** и **понятность** для пользователя. ***Интерфейс не должен быть перегружен*** во избежание утомляемости пользователя***. Интерфейс не должен быть слишком сложным***, чтобы пользователю не пришлось тратить много времени на его изучение. Эти требования являются актуальными и в случае разработки речевого интерфейса. Кроме того, учитывая, что исторически сложилось так, что графический интерфейс компьютерных систем является в настоящее время доминирующим, разработчикам интеллектуальных систем с речевым интерфейсом следует отдельно продумывать моменты, связанные с целесообразностью использования речевой формы взаимодействия с компьютером. Иными словами, ***речевой интерфейс должен быть востребован и уместен***.

К основным классам задач речевого интерфейса следует отнести:

- **синтез речи** – эта задача включает в себя комплекс подзадач и заключается в обеспечении возможности произнесение речи компьютером на основе произвольного орфографического текста;
- **анализ и распознавание речи** – комплекс задач, включающих запись, оцифровку и анализ речи для распознавания полученного речевого сообщения компьютерной системой;
- **понимание** (интерпретация) **речи** – это комплекс задач, связанных с анализом смысла речевых сообщений и формированием реакции (ответа) компьютерной системы. Часто указанная задача является подзадачей задачи распознавания речи;
- **распознавание голоса** – комплекс задач, включающих анализ особенностей голоса говорящего с целью выявления каких-либо его индивидуальных (личностных) особенностей и качеств. Данный комплекс задач называют также верификацией и идентификацией речи;
- **компьютерное клонирование голоса и дикции** [4] (***Лобанов Б.М.2002ст-Комп_К_П_Г***) – это создание близкой копии, но не биологической, а компьютерной, и не всего существа в целом (в данном случае человека), а только одной из его интеллектуальных функций: чтение произвольного орфографического текста. При этом ставится задача максимально полного сохранения персональных акустических особенностей голоса, фонетических особенностей произношения и акцента, а также просодической (интонационной) индивидуальности речи (мелодика, ритмика, динамика).

Помимо перечисленных выше задач, входящих в группу задач разработки собственно речевого интерфейса, следует также отметить, что имеется ряд вспомогательных задач, решением которых занимаются научно-технические коллективы, разрабатывающие речевые системы. Это связано с тем, что задача реализации речевого интерфейса до сих пор не решена окончательно. Еще есть много вопросов, ответы на которые ищут многие научные коллективы как в нашей стране, так и за рубежом. К таким задачам, в частности, относятся следующие:

- исследование особенностей фонетического строения речи различных естественных языков;

- исследование особенностей интонационной окраски речи различных языков;
- выявление наборов параметров для описания речи, используемых как для синтеза речи, так и для ее распознавания;
- разработка новых методов синтеза речи;
- исследование различий речи разных дикторов и, в частности, мужского и женского голоса;
- разработка новых методов распознавания речи;
- поиск оптимальных путей передачи речи по каналам связи;
- разработка специальных шумоподавляющих микрофонов;
- разработка специальной аппаратуры для исследования характеристик речи;
- разработка новых методов оцифровки и оптимального сжатия речевого сигнала;
- разработка специальных звуковых карт, ориентированных на синтез и анализ речи;
- формирование баз данных с «образцами» речи различных дикторов с целью повышения естественности звучания синтезированной речи;
- исследование строения речевого тракта человека и особенностей образования звуков речи;
- исследование строения органов слуха человека;
- исследование особенностей восприятия речи человеком;
- поиск путей оптимального использования речевого интерфейса в различных технических и бытовых системах и разработка соответствующих технологий и др.

Решение указанных задач выливается в достаточно долгую и кропотливую работу, но получаемые при этом результаты дают возможность надеяться на то, что в скором времени использование речевых интерфейсов станет более распространенным и доступным [5] (***Автом_Р_и_С_Р-2000сб***).

1.1. Исторический обзор проблемы распознавания и синтеза речи

Вопрос о возможности общения с технической системой интересовал человечество уже давно, с тех пор как начали появляться первые механические машины. Возникла идея научить машину говорить. Первые попытки создания в Росси синтезированной речи относятся к XVIII веку. Во времена правления Екатерины II Петербургская Академия Наук объявила конкурс на создание говорящей машины. Это был механический синтезатор речи, с помощью которого воспроизводились отдельные гласные звуки русской речи.

В XIX веке появление резонаторной теории Гельмгольца дало новый толчок в развитии речевых исследований. Речевой тракт человека рассматривался как последовательность резонаторов. Ученые пришли к выводу о том, что гласные звуки различаются резонансными частотами, названными впоследствии **формантами**.

Серьёзные исследования в области речи относятся к началу XX века. В 1939 г. американский учёный Дадли создал первый **вокодер**, который осуществлял запись, сжатие и воспроизведение речи.

Основными историческими этапами и направлениями развития рассматриваемой проблематики являются следующие:

- развитие теория дифференциальных признаков;
- появление акустической теории речеобразования [6] (***Фант Г.1964кн-Акуст_Т_Р***);
- 40-е годы XX века: получение первых результатов в распознавании изолированных русских гласных;
- создание в г. Бнро Института речи, основные цели которого заключались в разработке вокодеров, решении задач верификация голоса, распознавании ключевых слов;
- начало 1965 г. XX века – 1-я Всесоюзная школа-семинар по автоматическому распознаванию слуховых образов (АРСО), собиравшая в лучшие годы до 250 участников. Последнее АРСО-17 было проведено в 1992 г. За эти годы советскими исследователями были предложены признанный во всем мире ДП-метод распознавания речи (см. раздел 5), формантный метод синтеза русской речи по тексту (см. разделы 3 и 4) и экспертный метод распознавания сонограмм (динамических спектрограмм – см. разделы 3 и 5). Таким образом, была заложена основа перехода к новому этапу речевых исследований – решению задачи распознавания речи неограниченного словаря;
- 1967 г. XX века – разработка метода динамического программирования (см. раздел 5), что фактически явилось революцией в принятии решений при распознавании речи;

- разработка метода коэффициентов линейного предсказания (см. раздел 3) анализа речевого сигнала;
- развитие экспертно-лингвистического метода, основанного на использовании комплекса акустико-фонетических знаний;
- проект АРПА (США);
- появление метода скрытых марковских моделей для решения задачи распознавания речи;
- 80-е годы XX века – появление первых коммерческих систем синтеза и распознавания речи. МОНИИС – система автоматической обзвонки. Распознавание изолированных речевых команд малого словаря;
- 90-е годы XX века – многоязычные синтезаторы речи, распознавание больших словарей;
- разработка пишущей машинки с голоса – система ДРАГОН и др.;
- проекты по созданию систем автоматического перевода (английский, немецкий, японский языки). Цепочка: ввод РС – распознавание – понимание – перевод через английский язык – синтез – РС;
- развитие компьютерной телефонии.

Ниже кратко описаны основные этапы развития отечественной проблематики синтеза речи (http://www.ssrlab.com).

Первая, примитивная модель синтезатора русской речи ФОНЕМАФОН-1 “заговорила” в начале 70-х гг. XX века, и успех в её создании был связан прежде всего с разработкой принципов **формантного синтеза** речевых сигналов (см. также разделы 3 и 4). В дальнейшем появилась усовершенствованная модель формантного синтеза речевых сигналов, а затем были оптимизированы характеристики формантных фильтров синтезатора речи последовательного типа. Важную роль в создании промышленных синтезаторов речи сыграла разработка цифрового формантного синтезатора. Ещё долгое время формантный синтезатор играл ключевую роль в системах синтеза речи по тексту, пока в конце 80-х - начале 90-х гг. XX века не был предложен новый микроволновой метод синтеза речевых сигналов (***Лобанов Б.М.1991ст-Микро_С_Р***) (см. также разделы 3 и 4).

В процессе развития сменилось три поколения систем синтеза речи по тексту, в основу которых были положены три существенно различных подхода к синтезу фонетических характеристик речи: **фонемно-артикуляторно-формантный**, **фонемно-формантный** и **фонемно-микроволновой**. Толчком к появлению первого подхода послужило исследование коартикуляции на акустическом уровне, которое позволило осуществить текущее определение формантных частот по функциям движения артикуляторов (см. раздел 4). В результате была разработана модель артикуляторного синтеза речи по печатному тексту. Второй подход удалось реализовать благодаря развитию акустической теории коартикуляции и редукции (***Lobanov B.1982art-On_the_Acous***), созданию методики построения формантных портретов фонем, а также созданию алгоритмов синтеза формантных параметров и вычислению фонемных портретов для синтеза речи по тексту (***Аксютина И.В..1986ст-Алгор_В_Ф***). Третий подход сформировался в начале 90-х гг. XX века и получил название микроволнового синтеза речи по тексту [12] (***Лобанов Б.М.1991ст-Микро_С_Р_по_Т***).

Несмотря на исключительную важность просодических характеристик (см. раздел 2) для качественного синтеза речи, сведения о закономерностях их поведения в русской речи были крайне скудными. Поэтому в начале 70-х гг. XX века были проведены эксперименты по восприятию русской интонации односложной синтетической фразы, проведен анализ и синтез просодических характеристик двухсложного слова, разработаны правила синтеза просодических характеристик односинтагменных фраз и сформулированы принципы автоматического синтеза интонационных структур. В дальнейшем алгоритмы автоматического синтеза интонации по печатному тексту были усовершенствованы. Это касалось алгоритмов синтеза по тексту мелодического и ритмического контуров, а также моделей синтеза мелодического контура русских и английских фраз (***Карневская Е.Б..1982ст-Модел_С_М_К***). Были разработаны алгоритмы интонирования текста и многофакторная модель ритмики.

Разработанные методы синтеза речевого сигнала, а также методы синтеза фонетических и просодических характеристик речи позволили приступить к созданию целостных моделей синтеза речи по тексту. Первой такой моделью стал формантный синтезатор речи по последовательности аллофонов. Был разработан преобразователь графема-фонема для синтеза речи по орфографическому тексту, и вместе с моделью артикуляторно-формантного синтеза речи он стал основой для устройства синтеза речи. Были заложены также лингвоакустические основы двуязычного синтеза речи (***Карневская Е.Б..1980ст-Лингв_О; Lobanov B.1981art-Artic-F_S_S***) и разработан алгоритм синтеза многоязычной речи по тексту.

1.2. Речевой интерфейс: компьютерные системы распознавания и синтеза речи

Исходя из перечисленных выше задач, общая структура **речевого интерфейса** (РИ) включает два основных компонента:

- синтез речи;
- распознавание речи.

Так как каждая из задач речевого интерфейса является достаточно сложной, то в соответствие указанным компонентам ставятся два отдельных класса систем:

- системы автоматического синтеза речи;
- системы автоматического распознавания речи.

Каждый из указанных классов систем требует отдельного подробного рассмотрения и, кроме того, они могут быть реализованы и использованы отдельно друг от друга. Существуют, например, системы синтеза речи, назначение которых заключается только в озвучивании текстов, без необходимости распознавания. Эти системы постепенно внедряются в различные приложения в виде дополнительных сервисных функций. Иногда системы, использующие речевой интерфейс, или подсистемы, реализующие речевой интерфейс в составе другой более сложной системы, называют **речевыми системами**.

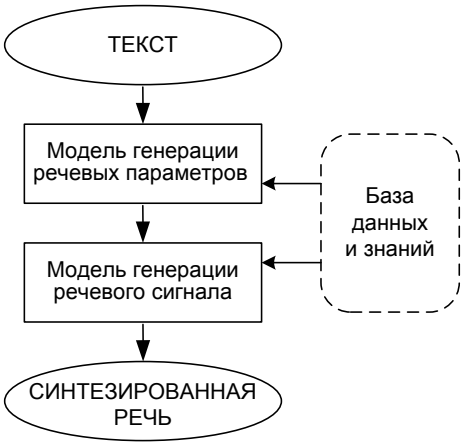
Ниже кратко рассматриваются основные особенности указанных систем. Более подробно отдельные вопросы их реализации будут рассмотрены в последующих разделах.

1.2.1. Системы автоматического синтеза речи

К л ю ч е в ы е п о н я т и я :
система автоматического синтеза речи, синтезатор речи, модель генерации речевых параметров, модель генерации речевого сигнала, синтезатор речевых параметров, синтезатор речевого сигнала, оценка качества синтеза речи, разборчивость, фразовая разборчивость, словесная разборчивость, слоговая разборчивость, звуковая разборчивость, естественность (натуральность) **речи, мультимодальность речи, многоязычие.**

О п р е д е л е н и е 1 . 1 . Под **системами автоматического синтеза речи** (иначе их еще называют **синтезаторами речи**) понимают системы, преобразующие орфографический текст и другую информацию в звучащую речь. Общепринятое в английской литературе обозначение – TTS (Text To Speech) System – системы преобразования текста в речь.

Упрощенная структурная схема системы автоматического синтеза речи представлена на рис. 1.2.



Р и с у н о к 1 . 2 . Структурная схема системы автоматического синтеза речи

О п р е д е л е н и е 1 . 2 . Под **моделью генерации речевых параметров** понимается блок преобразования входного орфографического текста в последовательность параметров, с помощью которых можно описать речь. Это могут быть артикуляторные параметры, либо параметры, связанные с акустикой речи, либо другие параметры, набор которых определяется, исходя из того, какая информация заключена в речевом потоке и каким образом она описана. Подробнее об этом будет рассказано ниже, в разделе 4.

О п р е д е л е н и е 1 . 3 . Модель генерации речевого сигнала – это блок преобразования речевых параметров в речевой сигнал, который воспринимает пользователь системы. Данный блок сопряжен с динамиками и в некоторых реализациях синтезаторов речи представляет собой только соответствующую аппаратную часть речевого интерфейса, а в некоторых – аппаратно-программную.

Фактически система автоматического синтеза речи – это совокупность двух компонент, которые часто называют **синтезатором речевых параметров** и **синтезатором речевого сигнала**. Оба этих компонента реализуются не только как набор программ, но и используют некую базу данных и знаний, содержащую информацию об особенностях организации естественного языка и о закономерностях, которые следует учитывать при синтезе речи. Кроме того, синтезатор речевого сигнала имеет аппаратно-программную реализацию, так как для того чтобы мы услышали звук, необходимо наличие, как минимум, звуковой платы и динамиков, подключенных к компьютеру. Таким образом, на выходе мы получаем звучащую синтезированную речь.

Следует отметить, что если в распознавании речи имеется такая важная характеристика, как объём словаря, то при синтезе речи это не является актуальным. Современные системы синтеза речи ориентированы на неограниченный объём словаря, благодаря тому, что они построены не с учетом знаний языка, а с учетом знаний об особенностях строения речи. Таким образом, система автоматического синтеза речи «умеет читать» практически любой естественно-языковой текст. Системы, которые работают с ограниченным словарём, не являются системами синтеза речи. Это простые синтезаторы, а точнее – воспроизводители некоторых записанных выражений, т.е. это системы типа цифрового магнитофона с произвольной выборкой сообщений.

При разработке систем автоматического синтеза речи очень важным является вопрос **оценки качества синтеза речи**. В процессе оценки качества учитываются следующие основные характеристики:

- разборчивость речи;
- естественность (натуральность) речи;
- мультимодальность речи;
- многоязычие.

Рассмотрим указанные характеристики более подробно.

Основным критерием оценки качества синтеза речи является **разборчивость** синтезированной речи. Очевидно, что чем выше разборчивость речи, тем более высокого класса синтезатор. Существуют различные способы (типы) оценки разборчивости, основными из которых являются следующие:

- звуковая (не менее 75 %);
- слоговая (не менее 85 %);
- словесная (не менее 99 %);
- фразовая (98 – 99 %);
- смысловая.

Обычно разборчивость измеряется (оценивается) следующим образом. Речевая система синтезирует какие-то элементы, например, фразы: «на улице хорошая погода», «вдалеке послышался выстрел», т.е. различные неожиданные для слушателя фразы. Указанные фразы фиксирует группа слушателей, каждый из которых пытается разобрать (распознать), что сказала машина. Например, предоставляется возможность прослушать 100 фраз, и слушатели поняли 98 – 99 % – это очень хорошая **фразовая разборчивость**.

При оценке **словесной разборчивости** слушателям предлагаются отдельные слова, самые разные, но осмысленные, никак между собой не связанные, из разных областей. Слушатели записывают слова, которые они поняли (расслышали). Затем подсчитывается количество понятых слов и рассчитывается процент относительно общего количества предложенных слов. При словесной разборчивости хорошим считается результат не менее 99 %.

В этом случае оценки **слоговой разборчивости** произносятся бессмысленные слоги (например, «псу», «ваз», «дус», «гры» и т.д.). При этом используются специальные таблицы частотной встречаемости слогов, с учетом которых формируются тестовые последовательности, которые подаются на вход речевой системы. Слушателям опять-таки следует записать все понятые ими слоги. Затем подсчитывается, сколько слогов услышано правильно: если примерно 85 % и выше, то разборчивость считается достаточной.

Звуковая разборчивость оценивается по тем же бессмысленным слогам, но считается не число неправильных восприятий (если хотя бы один звук в слоге был воспринят неправильно, то слог уже неправильный), а считается число звуков, т.е. фонем, воспринятых неправильно. Поскольку слоги состоят из различных звуков, то считается, что 75 % правильно воспринятых звуков – это уже неплохо.

При оценке разборчивости используются также градации, т.е. более градуированные оценки – какая разборчивость считается отличной, хорошей, удовлетворительной, неудовлетворительной.

Следующей характеристикой, используемой для оценки качества синтезатора речи, является **естественность** (натуральность) **речи**. Это субъективная характеристика, которая оценивается слушателями на основании их личного восприятия речи. Натуральность синтезированной речи зависит от многих факторов, например, могут быть натуральные звуки, но ритмическая сторона речи может быть сильно испорчена. Иначе говоря, слушатель может слышать понятную и разборчивую речь, произносимую вполне естественным голосом, но интонация при этом какая-то неестественная, «роботная». Это может также проявляться в том, что машина путает или “съедает” ударения. Натуральность речи можно оценить, но нет объективных критериев – это только субъективное впечатление слушателя. Ведь даже разные люди имеют разное произношение, которое иногда может даже показаться неестественным. При реализации речевых синтезаторов интонация и ритмика речи определяются исходя из анализа входного предложения. Расстановка ударений в словах осуществляется в соответствии со словарём и с учётом синтаксических правил.

Под **мультимодальностью речи** понимают отражение эмоционального состояния говорящего, индивидуальность его голоса, стиль речи, акцент и т.п. В системах автоматического синтеза речи эта характеристика выражается в возможности синтеза различных типов голосов и их индивидуальных особенностей. Эта возможность относится к экстралингвистическим способностям системы, так как не связана с языковыми и собственно речевыми особенностями реализации. К мультимодальности речи, в частности, относят поддержку мужских и женских голосов, различные голосовые модуляции (бас, баритон). Сюда же относится возможность синтеза некоторых эмоциональных компонент, содержащихся в речи, таких, как волнение, гнев, ласка и т.п. Не всегда хорошо, когда синтезатор «говорит» безразлично и монотонно; во многих случаях полезно, когда он «говорит», моделируя эмоции человека.

В отличие от мультимодальности, **многоязычие** относится к лингвистическим способностям и подразумевает возможность синтеза речи на нескольких естественных языках. Например, на русском, белорусском, английском и т.д.

Более подробно особенности разработки систем автоматического синтеза речи рассмотрены в разделе 4.

В настоящее время системы автоматического синтеза уже получили довольно широкое распространение (***Кучеров В.Я..1983кн-Синте_Р***). Примерами применения систем синтеза речи являются так называемые экранные чтецы, телефонные информаторы. Более подробно примеры применения систем синтеза речи рассмотрены в подразделе 1.3.

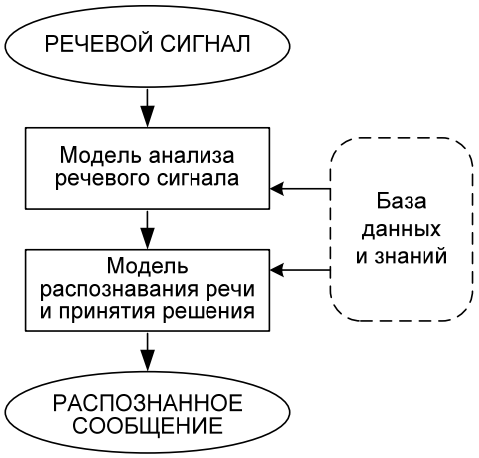
1.2.2. Системы автоматического распознавания речи

Ключевые понятия:
система автоматического распознавания речи, модель анализа речевого сигнала, модель распознавания речи и принятия решения, точность распознавания, система распознавания слитной речи, дикторонезависимое распознавание речи.

Определение 1.4. Под **системами автоматического распознавания речи** (САРР) понимают системы, преобразующие входную речь (речевой сигнал) в распознанное сообщение. При этом распознанное сообщение может быть представлено как в форме текста этого сообщения, так и преобразовано сразу в форму, удобную для его дальнейшей обработки с целью формирования

ответной реакции системы. Изначально перед системой автоматического распознавания речи ставится задача преобразования текста в речь. Поэтому в английской литературе эти системы называются Speech To Text System. Часто системы автоматического распознавания речи называют также просто системами распознавания речи (СРР).

Упрощенная структурная схема системы автоматического распознавания речи приведена на рис.1.3.



Р и с у н о к 1 . 3 . Структурная схема системы автоматического распознавания речи

О п р е д е л е н и е 1 . 5 . Под **моделью анализа речевого сигнала** понимают блок, в задачи которого входит анализ входного сигнала, во-первых, с целью отнесения его к числу речевых, а во-вторых, для выделения в составе полученного сигнала компонент, которые являются основными для распознавания полученного сообщения. К таким компонентам относятся параметры, описывающие речь, аналогичные тем, которые формируются в процессе синтеза речи. Набор указанных параметров зависит от избранного метода распознавания, что более подробно будет рассмотрено ниже, в разделе 5.

О п р е д е л е н и е 1 . 6 . Модель распознавания речи и принятия решения – это блок, в рамках которого осуществляется формирование распознанного сообщения на основе анализа последовательности параметров, полученных из первого блока. Например, если используется формантная модель описания речи, то на основе полученных в первом блоке частот формант строится последовательность распознанных фонем, составляющих входное сообщение. При этом осуществляется принятие решения о том, распознано ли входное сообщение правильно. При принятии решения, в частности, возможны следующие решения: сообщение распознано правильно (подтверждением этого является текст, соответствующий нормам естественного языка) либо сообщение не распознано или распознано не правильно (такое решение принимается в случае наличия в распознанном сообщении явных, трудно исправимых автоматически ошибок или вообще полной бессмыслицы).

Очевидно, что главным показателем качества системы распознавания речи является точность распознавания. При этом на СРР накладываются разного рода ограничения, так как в целом задача распознавания речи является на порядок сложнее задачи синтеза. При распознавании речи очень многое зависит от условий распознавания, так как «уши» компьютера существенным образом отличаются от органов слуха человека. Для разработки высококачественной СРР не достаточно знаний только физических (акустических) особенностей речи. Необходимо также знание особенностей восприятия речи человеком и строения его органов слуха. Например, человек способен из услышанного им сообщения выявлять главное, что позволяет ему слышать даже в очень неблагоприятных условиях. Понимая смысл услышанного, человек может автоматически восполнять пробелы, т.е. догадываться о том, что ему не удалось расслышать. Комплекс указанных проблем исследуется в области биологии и психологии (в частности, в рамках когнитивной психологии [7] (***Андерсон Дж.2002кн-Когни_П***)).

В качестве ограничений, накладываемых на САРР, можно привести следующие характеризующие их параметры:

- вид распознаваемой речи (пословное произношение с паузами в стиле речевых команд; четкое произношение без пауз в стиле “диктант”; спонтанная речь);

- объём словаря (ограниченный до 100, 200 и т.д. слов; неограниченный);
- степень зависимости от диктора (дикторозависимые; дикторонезависимые);
- синтаксические ограничения (отдельные слова; типовые фразы; искусственный язык; естественный язык);
- условия приёма речевых сигналов (контактные микрофоны; удаленные на расстояние более 1 м микрофоны);
- условия применения СРР (слабые или сильные помехи);
- надежность распознавания.

О п р е д е л е н и е 1 . 7 . Надежность СРР (**точность распознавания**) связана с числом ошибок в распознавании слов и обычно определяется следующим образом:

$$E = \left(1 - \frac{S + I + D}{N}\right) \cdot 100\,\%,$$

где E – оценка надежности СРР;

$S,\ I,\ D$ – общее число замен, вставок и выпадения слов соответственно;

N – общее число слов в тексте.

При этом вставки и выпадения являются значимыми для **систем распознавания слитной речи**.

Особо следует выделить проблему реализации **дикторонезависимого распознавания речи**, сложность которой обусловлена многочисленными различиями в голосах дикторов. Очевидно, что дикторозависимые СРР не могут получить широкого распространения, так как рассчитаны только на одного диктора. Устранение отрицательного эффекта от различий голосов дикторов достигается путем адаптации СРР к конкретному пользователю. Проблема дикторонезависимости в СРР может быть решена одним из следующих способов:

- создание базы данных акустических эталонов многих дикторов;
- кластеризация дикторов по особенностям голоса в группы (кластеры);
- быстрая подстройка под голос конкретного диктора по ограниченному словарю “парольных” фраз (такие системы называют квазидикторонезависимыми СРР);
- текущая адаптация к голосу диктора в процессе работы СРР.

Более подробно особенности разработки систем автоматического распознавания речи рассмотрены в разделе 5.

В качестве примеров систем распознавания речи можно привести следующие: голосовая "мышь" компьютера, голосовой набор номера абонента, телефонные справочные системы, телефонные автосекретари. Более подробно примеры применения САРР рассмотрены в подразделе 1.3.

1.3. Практические приложения речевого интерфейса

К л ю ч е в ы е п о н я т и я :
мультимодальная система, мультимедийная система, диалоговая система, интеллектуальная система.

Как было сказано выше, использование речевого интерфейса является наиболее естественной и удобной формой взаимодействия человека с технической системой. В настоящее время уже существует ряд систем, которые используют речевой интерфейс, обеспечивая тем самым более оптимальное решение некоторых задач.

Прежде чем перейти к рассмотрению примеров практического использования речевого интерфейса, сравним его с наиболее распространенными в настоящее время средствами взаимодействия пользователя с компьютером: клавиатурой и дисплеем. Следует отметить по крайней мере три принципиальных отличия речевого интерфейса:

1) явный недостаток клавиатуры и дисплея заключается в том, что для общения с компьютером человеку нужно пройти специальную подготовку. В то же время речь – это естественный интерфейс для любого, даже неподготовленного человека. Речь снижает в резкой степени

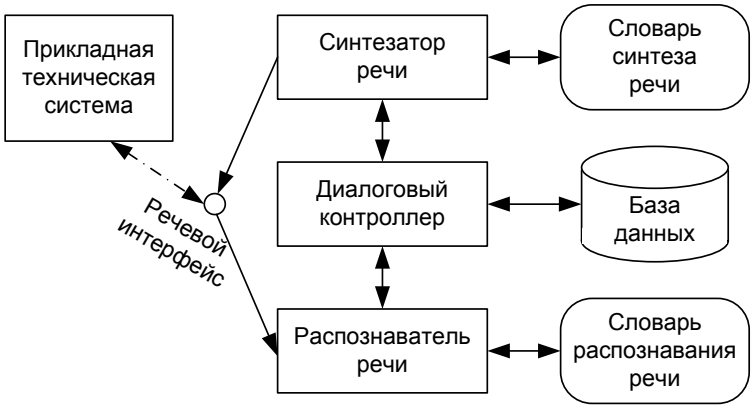
психологическое расстояние между человеком и компьютером. Если появляется речевой интерфейс, то круг пользователей компьютером может стать неограниченным;

2) речь сама по себе никак механически не привязана к компьютеру и может быть связана с ним через системы коммуникаций, например, телефон. Речевой интерфейс сокращает физическое расстояние между человеком и компьютером. Это дополнительно расширяет круг потенциальных пользователей компьютеров и делает речевой интерфейс идеальным средством для создания систем массового информационного обслуживания;
3) можно обращаться с компьютером в полной темноте, с закрытыми глазами, в условиях занятости рук рычагами управления, с завязанными руками и в другой экстремальной обстановке. Это свойство даёт оперативность и мобильность общения, освобождение рук и разгрузку зрительного канала восприятия при получении информации. Это исключительно важно, например, для диспетчера большой энергетической системы или пилота самолёта и водителя автомобиля. Кроме того, компьютерные системы становятся более доступными людям с нарушением зрения.

В настоящее время речевые компьютерные технологии уже достаточно широко распространены и развиваются в нескольких направлениях, основные из которых представлены на рис. 1.4.

Р и с у н о к 1 . 4 . Направления развития речевых компьютерных технологий

Итак, будущие машины – это **мультимодальные** и **мультимедийные** системы, т.е. такие системы, которые будут использовать, как и человек, различные каналы ввода и вывода информации. Следует также отметить, что в настоящее время практически любая современная компьютерная система должна представлять собой **диалоговую систему** (рис. 1.5), т.е. такую систему, пользователь которой мог бы с ней общаться как с равноправным коллегой по работе. Особенно это касается **интеллектуальных систем**.



Р и с у н о к 1 . 5 . Структура речевой диалоговой системы

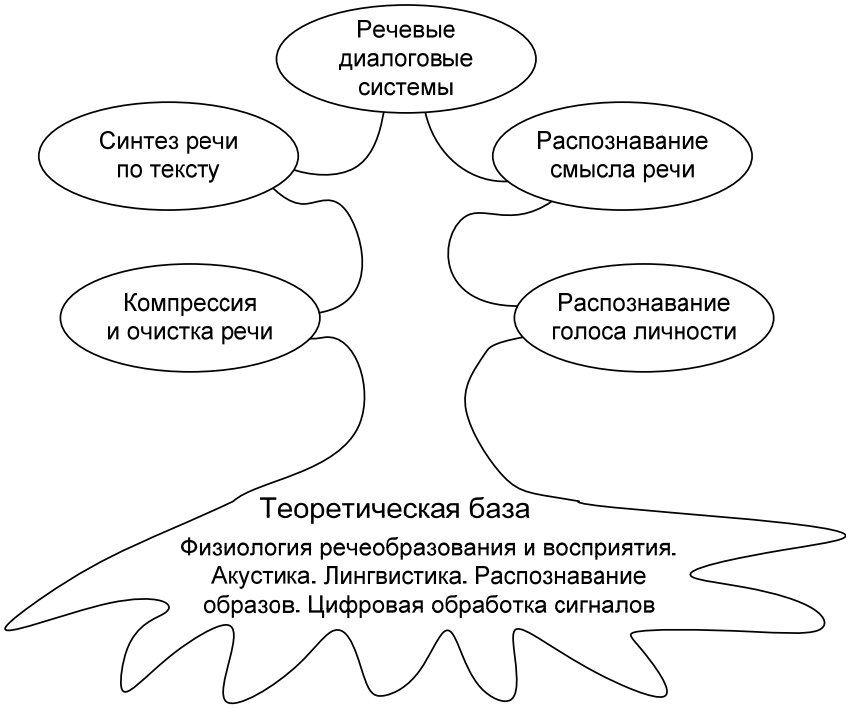
Вообще можно выделить следующие классы практических систем распознавания и синтеза речи:

1. **Синтезаторы речи**. Наибольший стимул их использования – это обслуживание слепых. Такие системы используют речевой дисплей, включая "подзвучивание" клавиатуры. С их помощью, например, можно осуществлять чтение электронной почты на расстоянии по телефону. В США

производится обслуживание клиентов по телефону с использованием кнопочного набора (tuch-tone) путём вывода синтезированного речевого сообщения.

2. **Системы распознавания речи**. В чистом виде – это "голосовая мышь" (Voice Mouse), для которой часть команд говорится голосом. Сравнительно недавно появились первые системы диктовки (Dictation Machines) – пишущие машины с голоса. Для них пока ещё существуют жесткие требования к манере пользователя говорить и требуется настройка на голос.
3. **Системы распознавания голоса**. Это системы, используемые преимущественно в сфере криминалистики, а также в системах защиты доступа.
4. **Диалоговые системы**. Это наиболее привлекательные системы, активно использующие и распознавание, и синтез речи, т.е. позволяющие вести диалог в форме речи. К ним относятся справочные центры (Call Centers), где кнопочный набор запроса заменяется на голосовой; интеллектуальные автоответчики (Phone Secretary), осуществляющие селекцию звонков, избирательную реакцию на звонки (в одних случаях – соединить по другому телефону, в других – послать сообщение на пейджер и т.д.).

Напомним также, что для реализации речевых систем необходимо использовать знания из нескольких предметных областей и решить ряд основных задач (рис. 1.6).



Р и с у н о к 1 . 6 . Основные виды речевых систем и теоретическая база для их реализации

Перечислим некоторые конкретные применения речевого интерфейса, которые уже существуют и используются в различных прикладных системах (***Кучеров В.Я..1983кн-Синте_Р***). К числу таких систем можно отнести следующие:

- системы поддержки безбумажных технологий: диктовка и формирование текстовых файлов на компьютере, системы подготовки документов, редакционно-издательские системы;
- речевые интерфейсы для пользователей-инвалидов по слуху и зрению;
- системы компьютерной телефонии (телефонные диалоговые информационно-справочные системы, включая справки по паролю, телефонные автосекретари, речевая электронная почта; речевой набор номера телефона и др.);
- системы речевого управления (информационные и навигационные системы, диспетчерские системы управления воздушным и наземным транспортом, тренажеры систем управления; интеллектуальные здания и др.);
- системы защиты доступа к базам данных, информации и объектам с использованием принципа парольных фраз ("Голосовой ключ");
- системы обнаружения голосовых сообщений (детекторы речи);

- системы защиты речевых сообщений (компрессия речи для повышения эффективности криптографической защиты речевых сообщений, повышение помехоустойчивости передачи речевых сообщений по сверхузкополосным каналам передачи данных и т.п.);
- системы-чтецы (например, система голосовых объявлений в общественном транспорте; системы голосового оповещения населения в чрезвычайных ситуациях);
- системы для криминалистической экспертизы на основе анализа голоса и речи;
- системы обучения языкам (в частности, иностранным), в число которых включаются также «говорящие» словари, речевые разговорники, системы обучения правильному произношению иностранных слов и т.п.;
- компьютерные системы обучения по различным предметным областям, использующие мультимодальный интерфейс;
- игровые компьютерные программы (в частности, компьютерные развивающие игры для детей).

1.4. Вопросы и задания на закрепление

1. Сформулируйте свое определение речевого интерфейса.

2. Перечислите и кратко охарактеризуйте основные задачи речевого интерфейса.

3. Представьте в виде схемы основные отличия речевого интерфейса от традиционного (графического) интерфейса.

4. Сформулируйте свое определение речевой системы. Сравните свое определение с определением, данным в тексте учебного пособия.

5. Представьте в виде схемы обобщенную структуру системы автоматического синтеза речи.

6. Как вы понимаете задачу автоматического синтеза речи?

7. Попробуйте перечислить основные научные направления, знания которых необходимо использовать при разработке систем автоматического синтеза речи.

8. Представьте в виде схемы обобщенную структуру системы автоматического распознавания речи.

9. Как вы понимаете задачу автоматического распознавания речи?

10. Попробуйте перечислить основные научные направления, знания которых необходимо использовать при разработке систем автоматического распознавания речи.

11. Сформулируйте свое определение мультимодальной системы.

12. Сформулируйте свое определение мультимедийной системы.

13. Перечислите основные сходства и отличия мультимодальных и мультимедийных систем.

14. Изобразите в виде схемы структуру диалоговой системы.

15. Перечислите основные направления использования речевых систем.

16. Приведите примеры использования речевых систем, известные вам из повседневной жизни.

17. Отметьте для себя вопросы, возникшие при изучении данного раздела и попытайтесь по оглавлению данного пособия отметить разделы, в которых возможно найти на них ответы.