**Перспективы речевого интерфейса**

Писать о речевом интерфейсе сложно. С одной стороны, тема абсолютно не нова, с другой- активное развитие и применение этой технологии только начина­ется (в который раз). С одной стороны, успели сформировать­ся устойчивые стереотипы и пре­дубеждения, с другой - несмот­ря на почти полвека настойчивых усилий не нашли разрешения вопросы, стояв­шие еще перед родоначальниками речевого ввода. Как бы то ни было, продолжат­ся поиски такого интерфейса, ко­торый устроил бы всех. Собственно говоря, это как раз то, к чему человечество всегда стремилось в общении с компьютером.

Исследователи недалеко про­двинулись за прошедшие десятки лет, что заставляет некоторых спе­циалистов крайне скептически от­носиться к самой возможности реализации речевого интерфейса в ближайшем будущем. Другие считают, что задача уже практи­чески решена. Впрочем, все зави­сит от того, что следует считать решением этой задачи.

Построение речевого интер­фейса распадается на три состав­ляющие.

1. Первая задача состоит в том, чтобы компьютер мог «понять» то, что ему говорит человек, то есть он доложен уметь извлекать из речи человека полезную ин­формацию. Пока что, на нынеш­нем этапе, эта задача сводится к тому, чтобы извлечь из речи смысловую ее часть, текст (пони­мание таких составляющих, как скажем, интонация, пока вообще не рассматривается). То есть эта задача сводится к замене клави­атуры микрофоном.
2. Вторая задача состоит в том, чтобы компьютер воспринял смысл сказанного. Пока речевое сообщение состоит из некоего стандартного набора понятных компьютеру команд (скажем, дуб­лирующих пункты меню), ничего сложного в ее реализации нет. Однако вряд ли такой подход бу­дет удобнее, чем ввод этих же ко­манд с клавиатуры или при помо­щи мыши. Пожалуй, даже удоб­нее просто щелкнуть мышкой по иконке приложения, чем четко выговаривать (к тому же мешая окружающим); «Старт! Главное меню! Бери!» В идеале компьютер должен четко «осмысливать» ес­тественную речь человека и пони­мать, что, к примеру, слова «Хва­тит!» и «Кончай работу!» означа­ют в одной ситуации разные по­нятия, а в другой - одно и то же.
3. Третья задача состоит в том, чтобы компьютер мог преобразо­вать информацию, с которой он оперирует, в речевое сообщение, понятное человеку.

Так вот, из этих трех задач достаточно ясное и окончатель­ное решение существует только для третьей. По сути, синтез речи - это чисто математическая за­дача, которая в настоящее время решена на довольно хорошем уровне. И в ближайшее время, скорее всего, будет совершен­ствоваться только ее техническая реализация.

Препятствием для окончательного решения первой задачи слу­жит то, что никто до сих пор тол­ком не знает, каким образом мож­но расчленить нашу речь, чтобы извлечь из нее те составляющие, в которых содержится смысл. В том звуковом потоке, который мы выдаем при разговоре, нельзя различить ни отдельных букв, ни слогов , об этом более подробно я Вам расскажу позже .. Во всяком случае, после предварительной тренировки современные системы распознавания речи работают довольно сносно и делают оши­бок не больше, чем делали оптические системы распознавания пе­чатных символов лет пять-семь назад.

Что касается второй задачи, то она, по мнению большинства спе­циалистов, не может быть решена без помощи систем искусственно­го интеллекта. Последние, как из­вестно, пока не созданы, хотя боль­шие надежды возлагаются на по­явление так называемых кванто­вых. Если же подобные устройства появятся, это будет оз­начать качественный переворот в вычислительных технологиях, и тогда, как знать, может быть , мно­гие теперешние подходы к рече­вому интерфейсу вообще окажут­ся ненужными.

Поэтому пока удел речевого интерфейса - всего лишь дубли­рование голосом команд, кото­рые могут быть введены с клави­атуры или при помощи мыши. А здесь его преимущества весьма сомнительны. Впрочем, есть одна область, которая для многих может ока­заться очень привлекательной. Это речевой ввод текстов в компью­тер. Действительно, чем стучать по клавиатуре, гораздо удобнее продиктовать все компьютеру, чтобы он записал услышанное в текстовый файл. Здесь вовсе не требуется, чтобы компьютер «ос­мысливал» услышанное, а задача перевода речи в текст более или менее решена. Недаром большин­ство выпускаемых ныне программ «речевого интерфейса» ориенти­рованы именно на ввод речи.

Хотя и здесь есть место для скепсиса. Если читать вслух, четко выговаривая слова, с паузами, монотонно, как это требуется для системы распознавания речи, то на машинописную страничку у меня уйдет пять минут. Печатаю на клавиатуре я с той же скоростью. Но сочиняю, при наличии вдох­новения, раза в два-три медлен­нее, а без оного - медленнее раз в пять, так что скорость «ввода» и при диктовке и при работе на кла­виатуре у меня абсолютно одинакова. Но вот сочинять и одновременно выговаривать сочиняемое с четкой артикуляцией , хоть убей, те не смогу.

Первый - и, пожалуй, основ­ной - вопрос касается области применения. Поиск приложений, где распознавание речи могло бы продемонстрировать все свои достоинства, вопреки устоявше­муся мнению, является задачей далеко не тривиальной. Сложив­шаяся практика применения ком­пьютеров вовсе не способствует широкому внедрению речевого интерфейса.

Для подачи команд, связан­ных с позиционированием в про­странстве, человек всегда поль­зовался и будет пользоваться жес­тами, то есть системой «руки- глаза». На этом принципе построен современный графический ин­терфейс. Перспектива замены клавиатуры и мыши блоком рас­познавания речи абсолютно от­падает. При этом выигрыш от воз­ложения на него части функций управления настолько мал, что не смог предоставить достаточных оснований даже для пробного внедрения в массовых компьюте­рах на протяжении уже более три­дцати лет. Именно таким сроком оценивается существование ком­мерчески применимых систем распознавания речи.

Для иллюстрации своих аргументов возможно, несколько спор­ных утверждений рассмотрю перспективу и основные пробле­мы применения систем речевого ввода текстов, особенно активно продвигаемых в последнее время.

Для сравнения: спонтанная речь произносится со средней ско­ростью 2,5 слов в секунду, про­фессиональная машинопись - 2 слова в секунду, непрофессио­нальная - 0,4. Таким образом, на первый взгляд, речевой ввод имеет значительное превосходство по производительности. Однако оценка средней скорости диктов­ки в реальных условиях снижается до 0,5-0,8 слова в секунду в связи с необходимостью четкого произ­несения слов при речевом вводе и достаточно высоким процентом ошибок распознавания, нуждаю­щихся в корректировке.

Речевой интерфейс естественен для человека и обеспечивает допол­нительное удобство при наборе тек­стов. Однако даже профессиональ­ного диктора может не обрадовать перспектива в течение нескольких часов диктовать малопонятливому и немому (к этому я еще вернусь) ком­пьютеру. Кроме того, имеющийся опыт эксплуатации подобных сис­тем свидетельствует о высокой веро­ятности заболевания голосовых связок операторов, что связано с неиз­бежной при диктовке компьютеру монотонностью речи.

Часто к достоинствам речевого ввода текста относят отсутствие не­обходимости в предварительном обучении. Однако одно из самых слабых мест современных систем распознавания речи- чувстви­тельность к четкости произноше­ния- приводит к потере этого, казалось бы, очевидного преиму­щества. Печатать на клавиатуре оператор учится в среднем 1-2 месяца. Постановка правильного произно­шения может занять несколько лет.

Существует и еще одно непри­ятное ограничение применимо­сти, сознательно не упоминаемое, на мой взгляд, создателями сис­тем речевого ввода. Оператор, взаимодействующий с компьютером через речевой интерфейс, вынужден работать в звука изолированном отдельном помещении либо пользоваться звукоизоли­рующим шлемом. Иначе он будет мешать работе своих соседей по офису, которые, в свою очередь, создавая дополнительный шумо­вой фон, будут значительно за­труднять работу речевого распо­знавателя. Таким образом, рече­вой интерфейс вступает в явное противоречие с современной ор­ганизационной структурой пред­приятий, ориентированных на коллективный труд. Ситуация не­сколько смягчается с развитием удаленных форм трудовой дея­тельности, однако еще достаточ­но долго самая естественная для человека производительная и по­тенциально массовая форма поль­зовательского интерфейса обре­чена на узкий круг применения.

Ограничения применимости систем распознавания речи в рам­ках наиболее популярных тради­ционных приложений заставляют сделать вывод о необходимости поиска потенциально перспектив­ных для внедрения речевого ин­терфейса приложений за преде­лами традиционной офисной сфе­ры, что подтверждается коммер­ческими успехами узкоспециали­зированных речевых систем. Са­мый успешный на сегодня проект коммерческого применения рас­познавания речи - телефонная сеть фирмы АТ&Т. Клиент может запросить одну из пяти категорий услуг, используя любые слова. Он говорит до тех пор, пока в его высказывании не встретится одно из пяти ключевых слов. Эта систе­ма в настоящее время обслужива­ет около миллиарда звонков в год.

Несмотря на то, что одним из наиболее перспективных направ­лений для внедрений систем рас­познавания речи может стать сфе­ра компьютерных игр, узкоспециа­лизированных реабилитационных программ для инвалидов, телефонных и информационных сис­тем, ведущие разработчики рече­вого распознавания наращивают усилия по достижению универса­лизации и увеличения объемов словаря даже в ущерб сокращению процедуры предварительной на­стройки на диктора.

Даже Билл Гейтс, являющий со­бой в некотором смысле идеал праг­матизма, оказался не свободен от исторически сложившихся стерео­типов. Начав в 95-96 году с разра­ботки собственной универсальной системы распознавания речи, он, окрыленный первыми и, пожалуй, сомнительными успехами, в 97-м провозгласил очередную эру по­всеместного внедрения речевого ин­терфейса. Средства речевого ввода планируется включить в стандарт­ную поставку новой версии Windows NT- чисто офисной операционной системы. При этом руко­водитель Microsoft упорно повто­ряет фразу о том, что скоро можно будет забыть о клавиатуре и мыши. Вероятно, он планирует продавать вместе с коробкой Windows NT аку­стические шлемы вроде тех, кото­рые используют военные летчики и пилоты «Формулы 1». Кроме того, неужели в ближайшем бу­дущем прекратится выпуск Word, Ехсеl и т. д. ? Управлять графическими объектами экрана голосом, не имея возможности помочь руками, бо­лее чем затруднительно.

Будущее речевого интерфейса в не меньшей степени зависит от умения современных исследова­телей и разработчиков не только создать технологическую основу речевого ввода, но и гармонично слить технологические находки в единую логически завершенную систему взаимодействия «чело­век-компьютер». Основная ра­бота еще впереди.

**Базовая технология**

Не следует путать термины «понимание» и распознавание» речи. В то время как второй непосредственно относится ктехноло­гии преобразования акустических речевых сигналов в последова­тельность символов машинной кодировки, например ASCII. первый подразумевает анализ более высоких уровней (прагматический, семантический и т. *д.)* и формирование на его основе представле­ния о смысловом содержании высказывания. Дальнейшее разграничение задан укрепилось благодаря коммерческому успеху узкоспециализированных систем, ни в малейшей степени не нуждающихся, напри­мер, в модуле анализа контекста высказывания.

Традиционно процесс распознавания речи подразделяется на несколько этапов. На первом - производится дискретизация непрерывного речевого сигнала. преобразованного в электрическую форму Обычно частота дискретизации составляет 10-11 кГц. разрядность- 8 бит, что считается оптимальным для работы со словарями небольшого объема (10-1000 слов) и соответствует качеству передачи речи телефонного канала (ЗГц- 3.4кГц). понятно что увеличение объема активного словаря должно со­провождаться повышением частоты оцифровки н в некоторых случаях *-* поднятием разрядности.

На втором этапе дискретный речевой сигнал подвергается очистке от шумов и преобразуется в более компактную форму. Сжатие производится посредством вычисления через каждые 10 мс некоторого набора числовых параметров (обычно не более 16) с минимальными потерями информации, описывающей данный речевой сигнал. Состав набора зависит от особенностей реализации системы. Начиная с 70-х годов наиболее популярным методом (практически стандартом) построения сжатого параметрического описания стало линейно-предиктивное кодирование (ЛПК), в основе которого лежит достаточно совершенная линейная модель голосового тракта. На втором месте по популярности находится, вероятно, спектральное описание, полученное с помощью дискретного преобразования Фурье.

Очень хорошие результаты, однако, могут быть достигнуты и при использовании других методов, часто менее требовательных к вычислительным ресурсам, например клипирования. В этом случае регистрируется количество изменений знака амплитуды речевого сигнала и временные интервалы между ними. Получаемая в резуль­тате последовательность значений, представляющих собой оценку длительностей периодов сохранения знака амплитудой, несмотря на кажущуюся примитивность метода, достаточно полно представляет различия между произносимыми звуками. На таком методе предоб­работки основана, в частности, система распознавания речи, разра­ботанная в конце 80-х в НИИ счетного машиностроения (Москва).

Временной (10 мс) интервал вычисления был определен и обос­нован экспериментально еще на заре развития технологии авто­матического распознавания речи. На этом интервале дискретный случайный процесс, представляющий оцифрованный речевой сиг­нал считается стационарным, то есть на таком временном интер­вале параметры голосового тракта значительно не изменяются.

Следующий этап- распознавание. Хранимые в памяти компь­ютера эталоны произношения по очереди сравниваются с текущим участком последовательности десяти миллисекундных векторов, описывающих входной речевой сигнал. В зависимости от степени совпадения выбирается лучший вариант и формируется гипотеза о содержании высказывания. Здесь мы сталкиваемся с очень суще­ственной проблемой - необходимостью нормализации сигнала по времени. Темп речи, длительность произношения отдельных слов и звуков даже для одного диктора варьируется в очень широких пределах. Таким образом, возможны значительные расхождения между отдельными участками хранимого эталона и теоретически совпадающим с ним входным сигналом за счет их временного рассогласования. Достаточно эффективно решать данную пробле­му позволяет разработанный в 70-х годах алгоритм динамического программирования и его разновидности (алгоритм Витерби). Осо­бенностью таких алгоритмов является возможность динамического сжатия и растяжения сигнала по временной оси непосредственно в процессе сравнения с эталоном. С начала 80-х все более широкое применение находят марковские модели, позволяющие на основе многоуровневого вероятностного подхода к описанию сигнала производить временную нормализацию и прогнозирование продолжений , что ускоряет процесс перебора эталонов и повышает надежность распознавания.

**Что такое распознавание речи?**

На первый взгляд, все очень просто: вы произносите фразу, на которую техническая система реагирует адекватно .На самом деле за столь простой идеей кроются огромные сложно­сти.

Почему же между постановкой задачи и ее решением лежит дис­танция огромного размера? Рас­познавание речи - молодая, раз­вивающаяся технология. Ее очер­тания пока зыбки и изменчивы. Поэтому в статье пока больше во­просов, чем ответов. Я попытаюсь немного рассказать о технологиях распознавания речи, и, надеюсь, вам будет интересно.

#### Немного о терминах

Начнем с главного термина. Что есть речь?

Говоря о речи, мы должны раз­личать такие понятия, как «речь», «звуковая речь», «звуковой сиг­нал», «сообщение», «текст».

В нашем случае, в приложении к задаче распознавания такие поня­тия, как «речь» и «звуковая речь» означают одно и то же - некое генерируемое человеком звуковое сообщение, которое может быть объективно зарегистрировано, измерено, сохранено, обработано и, что важно, воспроизведено при по­мощи приборов и алгоритмов. То есть речь может быть представлена в виде некоего речевого сигнала, который в свою очередь может ис­пользоваться для обратного вос­произведения речи. То есть можно поставить знак эквивалентности ме­жду звуковой речью и ее представ­лением в виде речевого сигнала*.* При этом под понятием «сообщение»может скрываться любая по­лезная для получателя информа­ция, а не только текст. Например, если интересоваться не словами а интонациями, то сообщением будут просодические нюансы речи. Что же касается распознавания речи, то в нашем случае задача сводится к извлечению из речи текста.

Но здесь мы сталкиваемся с од­ним противоречием. Текст, как из­вестно, состоит из букв, слов, пред­ложений, - то есть он дискретен. Речь же в нормальных условиях звучит слитно. Человеческая речь, в отличие от текста, вовсе не состоит из букв. Если мы запишем на магнитофонную ленту или на диск ком­пьютера звучание каждой отдель­ной буквы, а потом попробуем ском­поновать из этих звуков речь, у нас ничего не получится.

Люди уже довольно давно дога­дались о том, что элементарные звуки, из которых состоит речь, не эквивалентны буквам. Поэтому  придумали понятие фонемы для обозначения элементарных звуков речи. Хотя до сих пор специалисты никак не могут решить - сколько же всего различных фонем суще­ствует. Есть даже такой раздел лин­гвистики - фонетика. Большинство авторов даже для одного и того же языко­вого диалекта приводят разное ко­личество фонем. В русском языке по одним данным 43 фонемы, по другим - 64, по третьим - более сотни... Но так уж повелось, что есть миф о незыблемости понятия фонемы. И о том, что речевой сиг­нал состоит непосредственно из ку­сочков сигнала, каждый из которых является фонемой. К сожалению, все далеко не так просто.

Поначалу ученые рассматрива­ли речевой сигнал как набор неких универсалий, расположенных друг за другом на временной оси, и считали этими универсалиями фо­немы. Однако дальнейшие иссле­дования речевых сигналов ника­ких фонем не обнаружили.

Тогда одни исследователи спра­ведливо решили, что при генера­ции речевых сигналов наблюдает­ся коартикуляция, то есть взаимо­проникновение соседних звуков (мышцы лица, язык и челюсти об­ладают разной инерцией). Значит, речевой сигнал должен состоять не из фонем, а из аллофонов - комбинаций «слипшихся» фонем.

Другие исследователи, подобно физикам, атаковали идею элемен­тарности фонем и стали утвер­ждать, что фонемы надо поделить на еще более короткие кусочки или даже вообще отказаться от этого понятия и «расчленять» речевой сигнал как-то иначе. Так родились фоноиды и еще масса авторских названий элементарных звуков.

А дальше все многозначитель­но замолчали. Каждый принялся рассматривать речевой сигнал со своей позиции, сообщай об успе­хах весьма туманно. Последнее, весьма вероятно, можно объяснить желанием сохранить ноу-хау. Вот такая картина. Люди изобрели це­лую кучу претендентов на универ­сальность. Конечно, в их основу положено прежде всего человеческоеощущение звука. Возмож­но поэтому фонемы ничем не луч­ше букв. А фоноиды, аллофоны и прочая - лишь усовершенство­ванная версия звуковогоделения речи. Может быть, в них и есть какой-то смысл. Мы ведь услы­шим. А технически-то сигнал со­стоит не из наших, человеческих компонентов восприятия. Сигнал можно разложить, отфильтровать, как-то еще преобразовать. Задача не в этом. Необходимо найти некий эквивалент, построить модель ме­ханизма восприятия звуковречи. Большой интерес для ученых, работающих в облас­ти распознавания речи, представ­ляют различные разделы лингвистики, науки о языках. Возможно, удачный синтез достижений этих наук и теории обработки речевых сигналов приведут к ус­пешному созданию систем распознавания .

##### Главные трудности фонемного подхода

Темп речи варьируется в широ­ких пределах, часто в несколько раз. При этом различные звуки речи растягиваются или сжимают­ся не пропо-рционально. Напри­мер, гласные изменяются значи­тельно сильнее, чем полугласные и особенно смычные согласные. Для так называемых щелевых зву­ков есть свои закономерности. (По­лугласные - это звуки при гене­рации которых необходимо участие голосовых связок, как и для гласных звуков, но сами они в оби­ходе считаются согласными. На­пример, так обычно звучат «м», «н», «л» и «р». Смычные звуки образуются при резком смыкании и размыкании органов артикуляции. Например «б», «л», «д», «т». Образование щелевых звуков свя­зано с шипением и прочими эф­фектами турбулентности в органах артикуляции. Можно назвать «в», «ж», «с», а также «ш» и другие шипящие. В качестве примеров для простоты намеренно не приведе­ны звуки, не имеющие буквенных обозначений.) Эта свойство называется временной нестационарностью образцов речевого сигнала. Произнося одно и то же слово или фразу в разное время, под влиянием различных факторов (настроения, состояния здоровья и др.), мы генерируем заметно не совпадающие спектрально-вре­менные распределения энергии. Это справедливо даже для дваж­ды подряд произнесенного сло­ва. Намного сильнее этот эффект проявляется при сравнении спек­трограмм одной и той же фразы, произнесенной разными людьми. Обычно этот эффект называют спектральной нестационарной сетью образцов речевого сигнала (см. примеры спектрограмм). В Изменение темпа речи и четко­сти произношения является при­чиной коартикуляционной нестационарности, означающей изме­нение взаимовлияния соседних звуков от образца к образцу. Проблема кластеризации слит­ной речи. Из непрерывного рече­вого потока довольно непросто вы­делить какие-либо речевые еди­ницы. Многие звуки «слипаются» либо имеют нечеткие границы.

##### Многообразие видов

Существующие системы распо­знавания речи можно классифи­цировать по разным признакам.

По назначению:

1. командные системы
2. системы диктовки текста.

По потребительским качествам:

1. диктороориентированные (тре­нируемые на конкретного диктора)
2. дикторонезависимые (рискую предложить термин «омнивойс»)
3. распознающие отдельные слова
4. распознающие слитную речь.

По механизмам функциониро­вания:

1. простейшие (корреляционные) детекторы
2. экспертные системы с различ­ным способом формирования и обработки базы знаний
3. вероятностно-сетевые модели принятия решения, в том числе нейронные сети.

Довольно трудно выбрать удоб­ный показатель качества работы системы распознавания речи. Наи­более просто такой показатель ка­чества вводится для командных систем. При тестировании в слу­чайном порядке произносятся все возможные команды достаточно большое число раз. Подсчитыва­ется количество правильно распознанных команд и делится на об­щее количество произнесенных команд. В результате получается оценка вероятности правильного распознавания команды в задан­ной при эксперименте акустиче­ской обстановке. Для систем дик­товки похожий показатель качест­ва может вычисляться при диктов­ке некоторого тестового текста. Очевидно, что это не всегда удоб­ный показатель качества. В дейст­вительности мы сталкиваемся с са­мыми различными акустическими обстановками. Но как быть со сменой дикторов и сопутствующей ей тре­нировкой системы?

В качестве примера разрешите взять на рассмотрение вариант простейшей командной системы распознавания речи. Функционирование системы осно­вано на гипотезе о том, что спек­трально-временные характеристи­ки команд-слов для отдельно взя­того диктора изменяются слабо. Акустическая модель такой систе­мы представляет собой преобразователь из речевого сигналов спектрально-временную матрицу и мо­жет служить типичным примером изобретательского подхода. В са­мом простом случае команда ло­кализуется во времени по паузам в речевом сигнале. Лингвистический блок способен обнаружить огра­ниченное число команд плюс еще одну, которая означает все осталь­ные неизвестные системе слова. Как правило, лингвистическая мо­дель строится как алгоритм поиска максимума функционала от вход­ного образца и образцов всего «словарного запаса» системы. Часто это обычный двумерный коррелятор. Хотя выбор размер­ности пространства описания и его метрики может широко варьиро­ваться разработчиком.

Уже исходя из «конструкции» описанной системы понятно, что она представляет собой скорее игрушку, нежели полезный инструмент. В на­стоящее время на рынке представ­лено множество коммерческих сис­тем распознавания речи с гораздо большими возможностями:

* Voice Type Dictation , Voice Pilot , ViaVoice от IBM
* Voice Assist Creative от Techonology
* Listen for Windows от Verbex и многие другие.

Некоторые из них (например, ViaVoice) спо­собны, как заявляют разработчи­ки, вводить слитную речь.

Лингвистические блоки совре­менных систем реализуют слож­ную модель естественного языка. Иногда она основана на математи­ческом аппарате скрытых цепей Маркова, иногда использует по­следние достижения технологии нейронных сетей либо других ноу-хау. Устройство же акустических блоков подобных систем держится в строгом секрете. По некоторым признакам можно догадаться, что акустический блок некоторых сис­тем пытается моделировать естест­венный слуховой аппарат.

**Речевой вывод.**

Речевой вывод информации из компьютера- проблема не ме­нее важная, чем речевой ввод. Это вторая часть речевого интерфей­са, без которой разговор с компь­ютером не может состояться. Я имею в виду прочтение вслух тек­стовой информации, а не проиг­рывание заранее записанных зву­ковых файлов. То есть выдачу в речевой форме заранее не из­вестной информации.

Фактически, благодаря синтезу речи по тексту открывается еще один канал передачи данных от компьютера к человеку, анало­гичный тому, какой мы имеем бла­годаря монитору. Конечно, труд­новато было бы передать рисунок голосом. Но вот услышать элек­тронную почту или результат по­иска в базе данных в ряде случаев было бы довольно удобно, осо­бенно если в это время взгляд за­нят чем-либо другим. Например, придя утром на работу в офис, вы могли бы поправлять галстуку зер­кала или возвращать на место при­ческу (может быть, даже подкра­шивать ногти ) в то время как ком­пьютер будет  читать вслух по­следние известия или почту. Или. например, в середине рабочего дня он может привлечь ваше вни­мание сообщением, что прибли­жается время заранее назначен­ной деловой встречи.

С точки зрения пользователя, наиболее разумное решение про­блемы синтеза речи - это вклю­чение речевых функций (в перс­пективе - многоязычных, с воз­можностями перевода) в состав операционной системы. Компьютеры будут озву­чивать навигацию по меню, читать (дублировать голосом) экранные сообщения, каталоги файлов, и т. д. Важное замечанием пользо­ватель должен иметь достаточные возможности по настройке голоса компьютера, в частности, при же­лании, суметь выключить голос совсем.

Вышеупомянутые функции и сейчас были бы не лишними для лиц, имеющих проблемы со зре­нием. Для всех остальных они соз­дадут новое измерение удобства пользования компьютером и зна­чительно снизят нагрузку на нерв­ную систему и на зрение. По моему мнению, сейчас не стоит во­прос, нужны синтезаторы речи в персональных компьютерах или нет. Вопрос в другом - когда они будут установлены на каждом ком­пьютере. Осталось ждать, может быть, год или два.

###### Методы синтеза речи

Теперь, после оптимистического описания ближайшего будущего давайте обратимся собственно к тех­нологии синтеза речи. Рассмотрим какой-нибудь хотя бы минимально осмысленный текст, например, эту статью. Текст состоит из слов, раз­деленных пробелами и знаками  препинания. Произнесение слов зависит от их расположения в пред­ложении, а интонация фразы - от знаков препинания. Более того, довольно часто и от типа приме­няемой грамматической конструк­ции: в ряде случаев при произне­сении текста слышится явная пауза, хотя какие-либо знаки препи­нания отсутствуют. Наконец, про­изнесение зависит и от смысла сло­ва! Сравните, например, выбор од­ного из вариантов за'мок» или «замо'к» для одного и того же слова «замок».

Обобщенная функциональная система синтеза

Структура идеализированной сис­темы автоматического синтеза ре­чи может быть представлена блок- схемой, изображенной на рис.1.

# Ввод текста

Блоки лингвистической Определение Исправление

Обработки языка текста ошибок

Подготовка текста входного текста

к озвучиванию

Нормализация текста

Лингвистический анализ

## Формирование Фонемный транскриптор

Просодических Приведение фонем

характеристик к единицам синтеза

Озвучивание Формирование управляющей информации

Получение звукового сигнала

Звук

Она не описывает ни одну из суще­ствующих реально систем, но со­держит компоненты, которые мож­но обнаружить во многих системах.

### Модуль лингвистической обработки

Прежде всего, текст, подлежащий прочтению, поступает в модуль лингвистической обработки. В нем производится определение языка , а также отфильтровываются не подлежащие произнесению символы. В некоторых случаях ис­пользуются спелчекеры (модули исправления орфографических и пунктуационных ошибок). Затем происходит нормализация текста, то есть осуществляется разделе­ние введенного текста на слова и остальные последовательности символов.Все знаки пунктуации очень информатив­ны.

Для озвучивания цифр разра­батываются специальные подблоки. Преобразование цифр в по­следовательности слов является относительно легкой задачей, но цифры имеющие разное значение и функцию, про­износятся по-разному.

Лингвистический анализ

После процедуры нормализации каждому слову текста необходимо приписать сведения о его произношении, то есть превратить в **цепочку фонем** или, иначе говоря, создать его фо­немную транскрипцию. Во многих языках, в том числе и в русском, существуют достаточно регулярные правила чтения - правила со­ответствия между буквами и фоне­мами (звуками), которые, однако могут требовать предварительной **расстановки словесных ударений.** В английском языке правила чте­ния очень нерегулярны, и задача данного блока для английского синтеза тем самым усложняется. В любом случае при определении произношения имен собственных, заимствований, новых слов сокращений и аббревиатур возника­ют серьезные проблемы. Просто хранить транскрипцию для всех слов языка не представляется воз­можным из-за большого объема словаря и контекстных изменении произношения одного и того же слова во фразе.

Кроме того, следует корректно рассматривать случаи **графиче­ской омонимии:** одна и та же последовательность буквенных сим­волов в различных контекстах по­рой представляет два различных слова/словоформы и читается по- разному (ср. выше приведенный

пример слова «замок»). Часто удается решить проблему неод­нозначности такого рода путем грамматического анализа, однако иногда помогает только исполь­зование более широкой семанти­ческой информации.

Для языков с достаточно регу­лярными правилами чтения од­ним из продуктивных подходов к переводу слов в фонемы является система контекстных правил, пе­реводящих каждую букву/буква - сочетание в ту или иную фонему, то есть **автоматический фонем­ный транскриптор.** Однако чем больше в языке исключений из правил чтения, тем хуже работает этот метод. Стандартный способ улучшения произношения систе­мы состоит в занесении нескольких тысяч наиболее употребительных исключений в словарь. Аль­тернативное подходу «слово - буква-фонема» решение предпо­лагает **морфемный анализ** слова и перевод в фонемы морфов (то есть значимых частей слова: при­ставок, корней, суффиксов и окон­чаний). Однако в связи с разными пограничными явлениями на сты­ках морфов разложение на эти элементы представляет собой зна­чительные трудности. В то же вре­мя для языков с богатой морфо­логией, например, для русского. словарь морфов был бы компакт­нее. Морфемный анализ удобен еще и потому, что с его помощью можно определять принадлежность слов к частям речи, что очень важно для грамматического ана­лиза текста и задания его просодических характеристик. В английских системах синтеза морфем­ный анализ был реализован в сис­теме MiTalk**,** для которой процент ошибок транскриптора составляет *5%.*

Особую проблему для данного этапа обработки текста образуют имена собственные.

Формирование просодических характеристик

К просодическим характеристи­кам высказывания относятся его тональные, акцентные и ритмиче­ские характеристики. Их физиче­скими аналогами являются часто­та основного тона, энергия и дли­тельность. Таким обра­зом, от системы синтеза следует ожидать примерно того же, то есть, что она сможет понимать имею­щийся у нее на входе текст, ис­пользуя методы искусственного интеллекта. Однако этот уровень развития компьютерной техноло­гии еще не достигнут, и большин­ство современных систем автома­тического синтеза стараются корректно синтезировать речь с эмоционально нейтральной интона­цией. Между тем, даже эта задача на сегодняшний день представля­ется очень сложной .

**Методы озвучивания**

Теперь скажу несколько слов о наиболее распространенных ме­тодах озвучивания, то есть о мето­дах получения информации, управляющей параметрами соз­даваемого звукового сигнала, и способах формирования самого звукового сигнала.

Самое широкое разделение стратегий, применяемых при оз­вучивании речи, - это разделе­ние на подходы, которые направ­лены на построение действующей модели рече-производящей сис­темы человека, и подходы, где ставится задача смоделировать акустический сигнал как таковой. Первый подход известен под на­званием артикуляторного синте­за. Второй подход представляется на сегодняшний день более про­стым, поэтому он гораздо лучше изучен и практически более успе­шен. Внутри него выделяется два основных направления - формантный синтез по правилам и компилятивный синтез.

**Формантные синтезаторы** ис­пользуют возбуждающий сигнал, который проходит через цифро­вой фильтр, построенный на не­скольких резонансах, похожих на резонансы голосового тракта. Разделение возбуждающего сиг­нала и передаточной функции го­лосового тракта составляет основу классической акустической тео­рии речеобразования.

**Компилятивный синтез** осуще­ствляется путем склейки нужных единиц компиляции из имеюще­гося инвентаря. На этом принципе построено множество систем, использующих разные типы единиц и различные методы составления инвентаря. В таких системах необ­ходимо применять обработку сиг­нала для приведения частоты ос­новного тона, энергии и длитель­ности единиц к тем, которыми должна характеризоваться синтезируемая речь. Кроме того, требу­ется, чтобы алгоритм обработки сигнала сглаживал разрывы в формантией (и спектральной в целом) структуре на границах сегментов. В системах компилятивного синтеза применяются два разных типа ал­горитмов обработки сигнала: LP (сокр. англ. Linear Prediction - линейное предсказание) и **PSQLA** (сокр. англ. Pitch Synchronous Overlap and Add). LP-синтез осно­ван в значительной степени на аку­стической теории речеобразования, в отличие от PSOLA-синтеза, который действует путем простого разбиения звуковой волны, состав­ляющей единицу компиляции, на временные окна и их преобразо­вания. Алгоритмы PSOLA позво­ляют добиваться хорошего сохра­нения естественности звучания при модификации исходной звуковой волны.

**Наиболее распространенные** **системы синтеза (иностранные языки)**

Наиболее распространенными системами синтеза речи на сего­дня, очевидно, являются системы, поставляемые в комплекте со зву­ковыми платами. Если ваш ком­пьютер оснащен какой-либо из них, существует значительная ве­роятность того, что на нем уста­новлена система синтеза речи - увы, не русской, а английской ре­чи, точнее, ее американского ва­рианта. К большинству оригиналь­ных звуковых плат Sound Blaster прилагается система **Creative Text- Assist,** а вместе со звуковыми кар­тами других производителей час­то поставляется программа **Mono­logue** компании **FirsfByte.**

**TexAssist** представляет собой реализацию формантного синте­затора по правилам и базируется на системе **DECTalk,** разработан­ной корпорацией **Digital Equip­ment** при участии известного аме­риканского фонетиста **Денниса Клатта** (к сожалению, рано ушед­шего из жизни). DECTalk до сих пор остается своего рода стандар­том качества для синтеза речи аме­риканского варианта английско­го. Компания **Creative Technologies** предлагает разработчикам использовать **TextAssist** в своих программах**.**

Поддерживаемые операционные системы - MS Windows и Windows 95; для Windows NT существует вер­сия системы DECTalk. изначально создававшейся для Digital Unix. Новая версия TextAsslst, объявлен­ная фирмой Associative Computing, Inc. и разработанная с использо­ванием технологий DECtalll и Cre­ative, является в то же время мно­гоязычной системой синтеза, поддерживая английский, немец­кий, испанский и французский языки. Это обеспечивается преж­де всего использованием соот­ветствующих лингвистических мо­дулей. разработчик которых- фирма Lemout & Hausple Speech Produсts признанный лидер в поддержке многоязычных рече­вых технологий. В новой версии будет встроенный редактор сло­варя, а также специализирован­ное устройство TextReader с кно­почным управлением работой синтезатора в разных режима), чтения текста.

Программа **Monologue,** пред­назначенная для озвучивания тек­ста, находящегося в буфере обме­на MS Windows, использует сис­тему **ProVoice.** ProVoice- ком­пилятивный синтезатор с исполь­зованием оптимального выбора режима компрессии речи и со­хранения пограничных участков между звуками, разновидность TD-PSOLA. Рассчитан на амери­канский и британский английский, немецкий, французский, латино­американскую разновидность ис­панского и итальянский языки. Ин­вентарь сегментов компиляции - смешанной размерности: сегмен­ты- фонемы или аллофоны. Компания FirstByte позициониру­ет систему ProVoice и программ­ные продукты, основанные на ней, как приложения с низким потреба пением процессорного времени. FirstByte также предлагает рассчитанную на мощные компьютеры систему артикуляторного синтеза PrimoVox для использования в приложениях телефонии. Для разработчиков: Monologue Win32 поддерживает спецификацию MicrosoftSAPI.

**Синтезатор русской речи**

В качестве примера рассмот­рим разработку **«Говорящая мышь»** клуба голосовых техноло­гий научного парка МГУ.

В основе речевого синтеза ле­жит идея совмещения методов конкатенации и синтеза по пра­вилам**.** Метод конкатенации при адекватном наборе базовых эле­ментов компиляции обеспечивает качественное воспроизведение спектральных характеристик ре­чевого сигнала, а набор правил - возможность формирования ес­тественного интонационно-про­содического оформления выска­зываний. Существуют и другие ме­тоды синтеза, может быть, в пер­спективе более гибкие, подающие пока менее естественное озвучи­вание текста. Это, прежде всего параметрический (формантный'' синтез речи по правилам или на основе компиляции, развиваемый для ряда языков зарубежными ис­следователями. Однако для реа­лизации этого метода необходи­мы статистически представитель­ные **акустика-фонетические ба­зы данных** и соответствующая компьютерная технология, кото­рые пока доступны не всем.

**Инструментарий синтеза русской речи**

Упоминавшийся выше инструмен­тарий синтеза русской речи по тексту позволяет читать вспух **сме­шанные русско-английские тек­сты.** Инструментарий представляет собой набор динамических библиотек (DLL), в который входят модули русского и английского синтеза, словарь ударений рус­ского языка, модуль правил про­изнесения английских слов. На вход инструментария подается слово или предложение, подле­жащее произнесению, с выхода поступает звуковой файл в фор­мате WAV или VOX, записываемый в память или на жесткий диск.

**Что дальше?**

А дальше... С одной стороны, нужно не забывать, что речь - эта все-таки одно из проявлений выс­шей нервной деятельности челове­ка, и потому вряд пи в ближайшие несколько лет стоит ожидать появ­ления систем распознавания речи по эффективности и удобству срав­нимых с секретарем-машинисткой, печатающей «со слов». С другой стороны, в мире технологий все ме­няется очень быстро, и не известие, что сложнее: расслышать непри­нужденно сказанную фразу или ра­зыграть красивый эндшпиль...

Думаю не будет секретом то , что любой находящийся в этой аудитории человек если он болен машиной , если он фанатик врятли воспринимает её как неодушевлённый предмет , как мебель . Скорее в кучку железа под таинственным названием компьютер мы вкладываем душу ,вкладываем себя посредством непрерывного общения в прямом смысле этого слова . Лично я не раз замечала за собою безсознательные вещи : набивая текст, составляя программу ,инсталируя приложения я регулярно бросаю компьютеру нелестные отзывы о ней же.Типа: Чего ты еще хочешь Захлопнись , или Ну и кретин же ты. Наивно пологая что когда нибудь она меня всё-таки услышит и на реплику : «Какой же ты балбес» ласково произнесет питание компьютера отключено , потеря всех не сохраненных данных . Именно по-этому темой своего реферата я выбрала близкую мне : Речевые технологии .Хотя я назвала бы ее более лирично : Узнай меня по голосу . В своем выступлении я хотела бы осветить не только проблемы и перспективы развития речевого интерфейса, но по рассуждать о том нужен ли он вообще и ой как не скоро окупят надежды потери времени и денег.