Современные технологии распознавания речи

Леонович А. А.
leonal@mail.ru

Введение

С момента появления первых ЭВМ одним из наиболее важных вопросов развития компьютерной техники был процесс взаимодействия человека с машиной. Долгое время это было доступно только узким специалистам – технологи «общались» с машиной через посредника-программиста. Такая ситуация просуществовала вплоть до появления диалогового интерфейса, когда пользователь смог лично вводить с клавиатуры адресованную машине команду и получить осмысленный ответ. Дальнейшее появление графического интерфейса, в котором отпала необходимость в знании человеком каких-либо команд, привела к повсеместному распространению персональных компьютеров.

Однако человек всегда стремился к более универсальному и естественному способу взаимодействия с ЭВМ. Еще в эпоху перфокарт в научно-фантастических романах человек с компьютером разговаривал, как с равным себе. Тогда же были предприняты первые шаги по реализации речевого интерфейса [1]. В 1971 г. была начата разработка самого крупного проекта, когда-либо предпринимавшегося на то время в области распознавания речи, после того, как Advanced Research Project Agency (ARPA) министерства обороны США приняло 5-летний проект по созданию машин, которые позволяют «понимать» произносимые слитно предложения и объем словаря которых составлял 1000 слов. В конце 1976 г. было представлено несколько систем, одной из которых была HARPY. Эта система правильно понимала 95% произносимых пятью операторами предложений, используя словарь объемом 1011 слов и строго ограниченную грамматику предложений.

Современные системы распознавания речи

В настоящее время речевое распознавание находит все новые и новые области применения, начиная от приложений, осуществляющих преобразование речевой информации в текст и заканчивая бортовыми устройствами управления автомобилем. Все многообразие существующих систем распознавания речи можно условно разделить на следующие группы:

1. Программные ядра для аппаратных реализаций систем распознавания речи;
2. Наборы библиотек, утилит для разработки приложений, использующих речевое распознавание;
3. Независимые пользовательские приложения, осуществляющие речевое управление и/или преобразование речи в текст;
4. Специализированные приложения, использующие распознавание речи;
5. Устройства, выполняющие распознавание на аппаратном уровне;
6. Теоретические исследования и разработки.

Рассмотрим каждую из этих групп подробнее.

1. Программные ядра для аппаратных реализаций

В основе любой речевой технологии лежит так называемый «engine» или ядро программы – набор данных и правил, по которым осуществляется обработка данных. В зависимости от назначения этого ядра различают TTS и ASR engine. TTS (Text-to-Speech) engine предоставляет возможность синтеза речи по тексту, а ASR (Automatic Speech Recognition) engine – для распознавания речи.

Существует несколько крупных производителей, занимающихся созданием ASR ядер и среди них такие компании, как SPIRIT, Advanced Recognition Technologies, IBM.

Компания SPIRIT занимается созданием программных средств для цифровой телефонии, сжатия речи, идентификации говорящего, для технологий VoIP и GPS [2]. ASR engine от SPIRIT разработан для распознавания речевых команд и применяется в различных приложениях, таких как голосовое управление устройствами, голосовой набор в hands-free устройствах, ввод персональных идентификационных кодов (PIN) в системах безопасности. Данное ядро встраивается в любые DSP или RISC платформы и поставляется в виде объектного кода.

Корпорация IBM уже более 30 лет занимается вопросами автоматического распознавания речи и достигла в этой области больших успехов. Так компания ProVox Technologies на основе программного ядра ViaVoice® от IBM [3] создала систему для диктовки отчетов врачей-радиологов VoxReports [4]. По результатам тестирований, данная система с точностью 95-98% распознает слитную речь нормального темпа (до 180 слов в минуту) в независимости от диктора.

Однако словарь системы ограничен набором специфических медицинских терминов.

Opera Software договорилась с IBM об интеграции в браузеры Opera технологии распознавания речи Embedded ViaVoice [5]. Использование Embedded ViaVoice позволит пользователям управлять браузером не только с помощью мыши и клавиатуры, но и голосом.

Технология распознавания речи все больше применяется в средствах подвижной связи. Так компания Advanced Recognition Technologies создала систему smARTspeak NG, встраиваемую в мобильные телефоны [6]. Сейчас система smARTspeak NG применяется в бесклавиатурных телефонах от Siemens [7], телефонах Panasonic стандарта TDMA в США и других.

Sakrament ASR Engine – программная разработка белорусской компании «Сакрамент» [8], рассчитанная на применение в различных аппаратных системах и программных приложениях, использующих технологии распознавания речи. Заявленные характеристики: точность распознавания 95-98%; дикторонезависимость; языконезависимость; распознавание слитной речи в виде выражений и небольших предложений. Однако в данной системе нет возможности обучения – дополнительные словари создаются по заказу, самой компанией «Сакрамент».

2. Наборы библиотек для разработки приложений

С развитием речевых технологий и все большим внедрением мобильных устройств, возникла идея применения речевого управления при построении сетевых приложений. Для этого было необходимо разработать унифицированный стандарт для интеграции речевых технологий.

Один из открытых стандартов на основе XML-языка – VoiceXML (Voice eXtensible Markup Language), первая версия опубликована в мае 2000 г. международным консорциумом World Wide Web (W3 Consortium) – предназначен для разработки интерактивных голосовых приложений (Interactive Voice Response, IVR) управления медиаресурсами. Цель создания стандарта - привнесение всех преимуществ web-программирования в разработку IVR-приложений [9].

Однако интерес к многомодальным приложениям, сочетающим распознавание речи с другими формами ввода информации (при помощи клавиатуры, пера или набора цифровых кнопок) побудил ряд компаний, в том числе Microsoft, поддержать проект SALT Forum (Speech Application Language Tags - теги языка речевых приложений). И теперь вокруг SALT и VoiceXML консорциума W3C формируются два разных лагеря [10, 11]. До сих пор компании не могут прийти к единому мнению о выборе главного стандарта и сейчас оба направления развиваются в равной степени.

Различные компании занимаются разработкой пакетов для создания речевых приложений, так называемых Software Development Kit (SDK), поддерживающих тот или иной стандарт. Так компания Philips создала пакет Speech SDK. Данный пакет поддерживает спецификацию Voice XML и выполнен для связи с C/C++ API [12].

Компаниями CompTek [13] и Philips совместно был создан SpeechPearl — продукт, представляющий из себя набор программных модулей, библиотек и утилит для разработки систем распознавания речи с поддержкой русского языка для телефонных приложений. В июне 2004 г. в сервисе Телепат, обеспечивающем управление электронными кошельками WebMoney Transfer с помощью мобильных и городских телефонов, начала работу система распознавания речи, созданная на основе SpeechPearl. Данный сервис стал первой в России коммерческой системой массового обслуживания, в которой поддерживается функция распознавания речи.

С другой стороны, корпорация Microsoft распространяет свой продукт - Microsoft Speech SDK. Он содержит набор компонентов, описывающих соответствующий программный интерфейс Windows Speech API, документацию, исходные тексты программы-заготовки (ее достаточно дополнить только собственным алгоритмом распознавания), а также системы распознавания и преобразования текста в речь Microsoft Speech Recognition и Microsoft Text-to-Speech [14].

3. Независимые пользовательские приложения

В настоящее время рынок программных распознавателей речи представлен множеством приложений. Рассмотрим наиболее известные из них.

Dragon NaturallySpeaking Preferred фирмы Dragon Systems [15] – единственная программа, приблизившаяся к тому, чтобы соответствовать заявленным характеристикам. В целом он очень близко подходит к достижению заявленной безошибочности распознавания - 95%. Хотя пакет Dragon и уступает некоторым из конкурентов в том, что касается перемещения по экрану, правки и форматирования, он превосходит всех в главном - способности с первого раза правильно записывать произнесенные слова. Изначально данный пакет не работает с русским языком.

Компания М.С. Технолоджи [16] разработала программу «Микросервис» для управления функциями операционных систем Windows 98/Me/2000/XP и ввода текста в любой редактор. Программа поддерживает русский и английский языки и

содержит словарь порядка 10000 слов. Также создана упрощенная версия – «Микросервис» Light. Здесь объем словаря ограничен 300 словами и 100 командами. Компания 1С приобрела права на это ПО и выпускает его под названием «Диктограф».

Однако, по данным тестирований, «Микросервис» от М.С. Технолоджи показал неудовлетворительные результаты – 30-50% правильно распознанных слов и команд [17].

К сожалению, российский рынок программных средств распознавания речи представлен единичными разработками. Из всех программ, изначально разрабатываемых для русского языка, только ПО от белорусской компании «Сакрамент» может конкурировать по качеству распознавания с зарубежными аналогами.

4. Специализированные приложения

Распознавание речи может применяться не только для ввода текста или подачи команд, но и для более специфичных целей. Так компания «Центр Речевых Технологий» разрабатывает и производит программные продукты, технологии и образцы техники для подразделений МВД, ФСБ, МЮ, МЧС, МО, служб экстренной помощи, центров обработки вызовов и для других пользователей, в деятельности которых особое значение придается регистрации и обработке речевой информации [18].

Компанией созданы следующие приложения: «ИКАР Лаб» – инструментальный комплекс криминалистического исследования фонограмм речи, «Трал» – автоматизированный комплекс распознавания дикторов в фонограммах телефонных переговоров, «Территория» – автоматизированная система диагностики диалектов и акцентов русской устной речи.

Германский институт DFKI, занимающийся разработками в области искусственного интеллекта, разработал систему, названную Verbmobil, способную переводить разговорную речь с немецкого на английский или японский и обратно, непосредственно произнесенную в микрофон [19].

Система выполнена в виде независимого сервера Verbmobil Server. Благодаря этому, Verbmobil удалось связать с сетью мобильных телефонов стандарта GSM. Теперь разноязычные абоненты, подключившись к Verbmobil Server могут общаться друг с другом непосредственно, принимая уже переведенную речь, при этом Verbmobil автоматически настраивается на язык говорящего. По данным экспериментов, точность переводов составляет 90%, что было проверено на 25000 тестовых фразах.

5. Устройства, выполняющие распознавание на аппаратном уровне

Для использования функций речевого распознавания в различных устройствах, роботах, игрушках, разрабатываются аппаратные методы решения данной проблемы. Так американская компания Sensory Inc. разработала интегральную схему Voice Direct™ 364 осуществляющую дикторозависимое распознавание небольшого числа команд (около 60) после предварительного обучения [20]. Перед началом эксплуатации модуль необходимо обучить всем командам, используемым в работе. Команды сохраняются во внешнюю память в виде образов размером 128 байт. Во время работы, образ очередной команды сравнивается с эталонными из памяти в нейросетевом модуле и принимается решение о совпадении.

Тайваньская технологическая корпорация Primestar Technology Corporation разработала собственный чип VP-2025, предназначенный для речевого распознавания [21]. Данное устройство осуществляет распознавание с помощью нейросетевого метода.

Кроме того, американскими учеными принято решение создать специализированный микропроцессор для распознавания речи. Исследования в данном направлении будут проводиться сотрудниками Университета Карнеги-Меллон в Питсбурге (Пенсильвания) и Калифорнийского университета в Беркли. Ожидается, что новый микропроцессор появится в течение ближайших двух-трех лет. Причем эффективность распознавания речи таким чипом должна будет в 100-1000 раз превысить аналогичный показатель применяемых сегодня программно-аппаратных комплексов [22].

6. Теоретические исследования и разработки

Разработкой теоретической базы в области речевых технологий занимаются множество исследовательских групп по всему миру. В первую очередь это такие крупные корпорации как IBM, Intel, Microsoft, AT&T. Эти компании занимаются теорией распознавания уже не один десяток лет и являются законодателями в этой области.

Из всего разнообразия научных разработок подробно рассмотрим работы отечественных исследовательских групп.

В лаборатории автоматизированных систем массового обслуживания Института проблем управления РАН более 30 лет ведутся исследования в области речевого распознавания. Главным научным и практическим направлением деятельности лаборатории в настоящее время является применение компьютерного распознавания слитной речи в системах обслуживания населения с возможностью использования русского и других языков [23]. Разработаны математические модели для описания процессов в системах распознавания речи.

Институт системного анализа РАН [24] занимается работами в области распознавания речи, которые ориентированы на решение следующих задач: развитие теоретической базы, разработка и программная реализация методов автоматического анализа речевых сигналов в реальном времени, позволяющих повысить качество систем синтеза, распознавания и кодирования речи. Принципиальная новизна предложенных решений состоит в использовании островного нейросетевого анализа речевого сигнала в корреляции с выделением устойчивых признаков и применении фонологических и других «инженерных» знаний (то есть знаний, основанных на содержательном исследовании процесса произнесения или процесса восприятия) о тонкой структуре речевого сигнала.

Разработки «Истра-Софт» [25] в области речевых технологий включают в себя следующие основные направления: сжатие речевых файлов, распознавание речи, синтез речи по тексту, идентификация личности по голосу. Был разработан алгоритм выделения фонем из слитной речи в реальном времени. Алгоритм производит адаптивный анализ параметров звуковой информации и отделение параметров голосовой щели от параметров артикуляционного фильтра, выделяет параметры сигнала, которые воспринимаются как определенный звук (фонема), включая интонацию, описывает все измеренные параметры математически кратко.

С 1996 года компания «СТЭЛ - Компьютерные Системы» в сотрудничестве с ведущими специалистами филологического факультета МГУ им. М.В. Ломоносова, Вычислительного центра РАН и ряда других организаций выполняет проект по созданию прототипа дикторонезависимой системы распознавания русской речи [26]. С методологической точки зрения проект основан на применении современных методов обработки речевого сигнала и аппарата скрытых Марковских моделей для описания фонетических и семантико-синтаксических закономерностей русского языка.

Перспективы разработки систем распознавания речи

Как видно, технологии речевого распознавания нашли свое применение в различных областях. Однако в данной области множество проблем все еще остаются не решенными, многие идеи требуют дальнейшего развития. Так, программы, работающие с изолированными словами, достигли высокой точности в командных системах – в наиболее распространенных современных приложениях точность распознавания составляет в среднем 95-99% и зависит в основном от уровня шума. В то же время задача распознавания слитной речи в достаточной степени не решена, хотя в случае ограниченного словаря системы такого типа существуют (VoxReports [4] на ядре ViaVoice [3], Verbmobil) и показывают высокие результаты по точности. В настоящее время множество работ посвящено проблеме распознавания слитной речи (ИПУ РАН [23], «Истра-Софт» [25], IBM [3]), т.к. именно такой тип речевого взаимодействия считается наиболее перспективным.

Рис. 1 Классификация систем распознавания речи

Важнейшим этапом обработки речи в процессе распознавания, является выделение информативных признаков, однозначно характеризующих речевой сигнал. Существует некоторое число математических методов, анализирующих речевой спектр. Здесь самым широко используемым является преобразование Фурье, известное из теории цифровой

обработки сигналов. Данный математический аппарат хорошо себя зарекомендовал в данной области, имеется множество методик обработки сигналов, использующих в своей основе преобразование Фурье. Не смотря на это, постоянно ведутся работы по поиску иных путей параметризации речи. Одним из таких новых направлений, является вейвлет анализ, который стал применяться для исследования речевых сигналов сравнительно недавно. Теория данного метода сейчас развивается учеными всего мира, и многие исследователи возлагают большие надежды на использование инструмента вейвлет анализа для распознавания речи.

Если рассмотреть речевые распознаватели с позиции классификации по механизму функционирования, то подавляющая их часть относится к системам с вероятностно-сетевыми методами принятия решения о соответствии входного сигнала эталонному – это метод скрытого Марковского моделирования (СММ), метод динамического программирования и нейросетевой метод (рис. 1). Например, нейронные сети могут быть использованы для классификации характеристик речевого сигнала и принятия решения о принадлежности к той или иной группе эталонов [27]. Нейросеть обладает способностью к статистическому усреднению, т.е. решается проблема с вариативностью речи. Многие нейросетевые алгоритмы осуществляют параллельную обработку информации, т.е. одновременно работают все нейроны. Тем самым решается проблема со скоростью распознавания – обычно время работы нейросети составляет несколько итераций. Сейчас многие разработчики используют аппарат нейронных сетей для построения распознавателей [19, 24, 27].

Однако, если сравнить показатели современных систем распознавания с показателями систем времен начала зарождения это области науки, то можно сказать, что за прошедшие десятки лет исследователи недалеко продвинулись. Это заставляет некоторых специалистов сомневаться относительно возможности реализации речевого интерфейса в ближайшем будущем [28]. Другие считают, что задача уже практически решена. Большинство экспертов сходится во мнении, что для развития распознавания речи потребуется какое-то время. В рамках своего проекта «Super Human Speech Recognition» IBM надеется к 2010 году разработать коммерческие системы, преобразующие речь в печатный текст точнее, чем человек [29].

Литература

1. Методы автоматического распознавания речи: В 2-х книгах. Пер. с англ./Под ред. У. Ли. – М.: Мир, 1983. – Кн. 1. 328 с., ил.
2. http://www.spiritdsp.com
3. http://www.ibm.com/software/speech/
4. http://www.provox.com
5. http://www.opera.com
6. http://www.artcomp.com
7. http://www.xelibri.com
8. http://www.sakrament.com
9. http://www.w3.org/TR/voicexml20/
10. *Шварц Э.* Авторские права на пути Voice XML. // Computerworld, №36, 2001 г.
11. http://www.intel.com
12. http://www.philips.com/speechrecognition/
13. http://www.comptek.ru
14. http://www.microsoft.com/speech/
15. http://www.dragonsys.com
16. http://www.mstechnology.ru
17. http://art.bdk.com.ru/govor/
18. http://www.speechpro.ru
19. http://www.dfki.de/verbmobil/
20. http://www.sensoryinc.com
21. http://www.ptmc.com.tw
22. http://www.cmu.edu
23. http://www.ipu.ru
24. http://www.isa.ru
25. http://www.istrasoft.ru
26. http://www.stel.ru/speech/frame.html
27. J.P. Hosom, R. Cole, and M. Fanty. Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding. //Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, July 1999.
28. *Чекмарев А.* Речевые технологии – проблемы и перспективы. // Компьютерра, №49 с. 26-43, 1997 г.
29. *Broersma M.* Speech recognition begins to makes itself heard. // news.zdnet.co.uk, October 2003.