УДК 004.934

РАСПОЗНАВАНИЕ РЕЧИ: ЭТАПЫ РАЗВИТИЯ, СОВРЕМЕННЫЕ ТЕХНОЛОГИИ И ПЕРСПЕКТИВЫ ИХ ПРИМЕНЕНИЯ

М.Ф. Бондаренко[1], А.В. Работягов[2], С.В. Щепковский[3]

[1] ХНУРЭ, г. Харьков, Украина,
[2] ХНУРЭ, г. Харьков, beloswet@kture.kharkov.ua Украина,
[3] ХНУРЭ, г. Харьков, Украина, svserg@kture.kharkov.ua

В статье проведен обзор развития систем распознавания речи, рассмотрены общие принципы их построения, перечислены актуальные проблемы этого направления. Также рассмотрены актуальные области применения и перспективы развития систем распознавания речи.

МЕТОДЫ РАСПОЗНАВАНИЯ РЕЧИ, СОВРЕМЕННЫЕ И ПЕРСПЕКТИВНЫЕ РЕЧЕВЫЕ ТЕХНОЛОГИИ

Введение

В настоящее время существуют многочисленные технические средства, могущие воспринимать (распознавать) произносимые речевые сообщения: компьютеры, медицинское электронное оборудование, автомобили, мобильные телефоны и др.

Что такое распознавание речи? На первый взгляд, все кажется очень просто: человек произносит слово (фразу), а техническая система адекватно реагирует на него: либо выполняет команду, содержащуюся в слове (фразе), либо набирает диктуемый текст, либо как-то иначе "распоряжается" извлеченной из фразы информацией.

Бурное развитие распознавания речи с помощью персонального компьютера (ПК) началось с 1993 г.

Две ключевых задачи распознавания речи – достижение 100 % распознавания на ограниченном наборе команд хотя бы для одного диктора и независимое от диктора распознавание непрерывного речевого потока в реальном масштабе времени произвольного языка с приемлемым качеством – до сих пор не решены, несмотря на многочисленные попытки решения этих задач в течение последних 50-ти лет.

Современные системы распознавания речи уже дают возможность пользователям диктовать слова (фразы) в обычной разговорной манере. Однако процесс непрерывного распознавания речи, дающий до 95 % качества распознавания при оптимальных условиях, все-таки дает на 100 знаков 5 ошибок. Около 200 ошибок на странице формата А4 – слишком много для профессиональной работы. Рассмотрим ставшую традиционной последовательность действий для компьютерного распознавания речевого сигнала.

Как правило, система распознавания речи состоит из двух моделей: акустической и лингвистической.

Компьютер записывает звук речи в виде цифрового сигнала и делит его на аудиофрагменты длительностью несколько миллисекунд. Акустическая модель отвечает за преобразование речевого сигнала в набор признаков, в которых отображена информация о содержании речевого сообщения. Программа выполняет сложный анализ речи, сравнивая аудиофрагменты с записанными в память речевыми образцами.

Лингвистическая модель анализирует информацию, получаемую от акустической модели, и формирует окончательный результат распознавания. На основе вероятностного расчета компьютер определяет, что именно мог произнести пользователь. В основе модели лежит понятие фонемы – наименьшей акустической единицы языка. В процессе обучения компьютер распознает наиболее важные признаки произношения пользователем фонем и записывает полученные данные в виде профиля пользователя. Для таких систем важно, чтобы в дальнейшем во время диктовки пользователь, по возможности, выдерживал мелодию речи и произношение.

1. Этапы развития систем распознавания речи

Создание устройств, способных воспринимать и "понимать" звучащую речь, имеет более короткую историю, чем построение "говорящих машин", синтезирующих речь. Следующие даты можно назвать основными вехами в развитии компьютерного распознавания речи.

1962 г. Первое коммерческое устройство речевого вывода: модель 7772 от IBM.

1984 г. Первая система распознавания речи на базе ЭВМ. На распознавание слова уходили минуты. Система различала примерно 5000 слов.

1986 г. Опытный образец системы речевого ввода Tangora 4. Благодаря специальному микропроцессору впервые стала возможна обработка речи на рабочем месте в реальном времени. В системе уже появилась функция контроля контекста.

1990 г. Dragon System представила первую американскую версию программы речевого ввода Dragon Dictate System.

1992 г. Технология Tangora в модели клиент-сервер. Используется RISC-система IBM RS/6000. Речевой ввод с ПК под OS/2.

1993 г. Появилась первая система речевого ввода для ПК – Personal Dictation от IBM; стоимость $1000. Одновременно выходит Philips Dictation System – первая система непрерывного распознавания речи.

1995 г. IBM представила на CeBIT систему диктовки VoiceType со специализированными словарями для медиков и адвокатов.

1997 г. Появилась система клиент-сервер Speech Magic от Philips. Lernout & Hauspie представила первую англоязычную систему распознавания речи.

2001 г. Microsoft выпускает комплект офисных приложений Office XP с поддержкой речевого ввода и управления.

Первые попытки в данной области относятся к 40-м годам прошлого века, и связаны они с появлением спектральных анализаторов – электрических устройств, позволяющих анализировать спектральные характеристики речевых сигналов. В СССР в это время было создано первое техническое устройство, которое могло распознавать гласные русского языка на основе разности энергии в 14 частотных полосах [1].

Развитие области знаний, связанной с анализом и распознаванием речевого сигнала, началось с решения задач передачи речи по узкополосным каналам связи с полосой пропускания меньшей, чем у обычной телефонной линии. Решение этой задачи привело к созданию вокодеров – устройств, выполняющих сокращение частотной полосы речевых сигналов для линий дальней связи. Первым успехом в данной области считается полосный вокодер американского инженера-связиста Х. Дадли [2]. Он представлял собой параметрический вокодер, фильтровавший спектр речи с интервалом в 20-30 мс на несколько полос, в каждой из которых измерялась энергия. Вокодер сначала осуществляет спектрально-временной анализ речевого сигнала, выделяя его акустические параметры, а затем может восстановить (ресинтезировать) исходный речевой сигнал на основании выделенных параметров. В отличие от предшествующих синтезаторов, вокодер Дадли был основан не на имитации артикуляции, а на воспроизведении акустических параметров речевого сигнала.

Серьезные работы по распознаванию речи начались в основном после Второй мировой войны. Первое устройство для распознавания речи появилось в 1952 г., оно могло распознавать произнесённые человеком цифры [3]. В AT&T Bell Labs была создана система распознавания отдельных цифр с помощью простого согласования акустических характеристик с шаблонами. Она представляла собой довольно примитивную систему, которая могла распознавать цифры, переданные голосом по телефону.

Для дальнейшего развития автоматического распознавания речи (АРР), большое значение имели метод динамической спектрографии (типа "Видимая речь") и широкое использование соответствующей аппаратуры в фонетических исследованиях. К концу 50-х годов на материале самых разных языков был накоплен большой исследовательский материал, который свидетельствовал о сложной природе соответствия между привычными для лингвистов представлениями речевых отрезков в виде последовательности фонем или аллофонов и физической реальностью звучащей речи. В начале 60-х годов компания IBM разработала и продемонстрировала "Shoebox" – предшественника современных систем распознавания речи. Это новаторское устройство распознавало и реагировало на 16 произносимых слов, включая цифры от 0 до 9. Оно было показано по телевидению и в павильоне IBM на мировой ярмарке 1962 г. в Сиэтле.

Достижения в области анализа и передачи речевого сигнала впервые в нашей стране были широко изложены в монографии М. А. Сапожкова "Речевой сигнал в кибернетике и связи" в 1963 г. Позже вышла работа большого коллектива авторов "Вокодерная телефония. Методы и проблемы" под редакцией А. А. Пирогова [4]. За рубежом методы анализа речевого сигнала были опубликованы Дж. Фланаганом в своей монографии немного позже М. А. Сапожкова.

Система распознавания на основе вероятностного подхода была создана Фраем и Денесом в лондонском University College. В этой системе впервые использовались вероятности переходов между фонемами. Начиная с 1971 г. Агентство перспективных исследовательских программ (DARPA) Министерства обороны США финансировало четыре конкурирующих пятилетних проекта по разработке высокоэффективных систем распознавания речи. Победителем этой программы и единственной системой, соответствующей требованиям по распознаванию словаря из 1000 слов с точностью 90%, стала система HARPY, разработанная в университете CMU. Окончательная версия этой системы была создана на основе системы Dragon, разработанной аспирантом того же университета Дж. Бейкером [5]. В этой системе для вероятностного моделирования слов речи впервые были использованы скрытые марковские модели [6]. Скрытая марковская модель является на сегодняшний день наиболее широко применяемым и эффективным подходом к проблеме построения акустической модели.

Почти одновременно с системой Dragon в компании IBM была разработана еще одна система на основе скрытых марковских моделей. Начиная с этих двух разработок, вероятностные методы в целом и скрытые марковские модели в частности стали доминировать в исследованиях и разработках по распознаванию речи [7, 8]. Использование данного подхода, ввиду своей эффективности, стало в настоящее время почти промышленным стандартом.

2. Возможности современных технологий

Увеличение вычислительных мощностей мобильных устройств позволило и для них создать программы с функцией распознавания речи. Среди таких программ стоит отметить приложение Microsoft Voice Command, которое позволяет работать со многими приложениями при помощи голоса. Еще одной интересной программой является Speereo Voice Translator, голосовой переводчик. SVT способна распознавать фразы, произнесенные на английском языке, и "говорить" в ответ перевод на одном из выбранных языков.

Интеллектуальные речевые решения, позволяющие автоматически синтезировать и распознавать речевой сигнал, являются следующей ступенью развития интерактивных голосовых систем (IVR). Использование интерактивного телефонного приложения в настоящее время не веяние моды, а жизненная необходимость. Снижение нагрузки на операторов контакт-центров и секретарей, сокращение расходов на оплату труда и повышение производительности систем обслуживания — вот только некоторые преимущества, доказывающие целесообразность подобных решений.

Таким образом, в телефонных интерактивных приложениях все чаще стали использоваться системы автоматического распознавания и синтеза речи. При этом системы распознавания являются независимыми от дикторов, то есть распознают голос любого человека.

Следующим шагом технологий распознавания речи можно считать развитие так называемых Silent Speech Interfaces (SSI) (Интерфейсов Безмолвного Доступа). Эти системы обработки речи базируются на получении и обработке речевых сигналов на ранней стадии артикулирования.

В настоящее время, каждый человек, разговаривая по сотовому телефону, пользуется т.н. липредерами – вокодерами, работающими на основе линейного предсказания речевого сигнала, используемыми в стандарте GSM. Однако до сих пор в области вокодерной связи не решена задача максимального сжатия речевого сигнала до фонемного уровня и передачи его с наименьшей скоростью 60 бит/с, что соответствует письменной передачи речи произносимой со средней для человека скоростью 10 фонем в секунду. Решение этой задачи непосредственно связано с распознаванием непрерывной звучащей речи.

В настоящее время на рынке представлено множество коммерческих систем распознавания речи:

– Voice Type Dictation, Voice Pilot и ViaVoice от IBM;

– Dragon Dictate и Naturally Speaking от Nuance Communications;

– Voice Assist от Creative Technology;

– Listen for Windows от Verbex и многие другие.

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

Компания Nuance Communications, в частности, постоянно обновляет свой программный продукт Dragon NaturallySpeaking, который позволяет надиктовывать текстовые документы, а также управлять работой компьютера с помощью голосовых команд. Нужно отметить, что данный инструмент распознавания достаточно хорошо работает только с разговорным английским.

Петербургская компания "Центр речевых технологий", целенаправленно занимающаяся технологиями распознавания речи, еще в 2008 г. создала технологию распознавания слитной русской речи "Руссограф", для создания которой был создан уникальный для России набор речевых баз данных, в который входят записи более чем 3000 дикторов общей длительностью около 300 часов, собранных с учетом 5 диалектных групп русского языка. Уникальность данной технологии заключается в том, что многочисленные системы распознавания речи, применяемые к другим языкам, не обеспечивают такого же качества распознавания при работе с русским языком. Сейчас эта технология развивается и адаптируется для применения в конечных программных продуктах.

3. Проблемы реализации систем распознавания речи

Рассмотрим аспекты, которые препятствуют глобальному решению проблемы качественного распознавания речи.

1. Темп речи варьируется в широких пределах, часто в несколько раз. При этом различные звуки речи растягиваются или сжимаются не пропорционально. 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", "п", "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.) Это свойство называется временной нестационарностью образцов речевого сигнала.

2. 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 нестационарностью образцов речевого сигнала (см. примеры спектрограмм).

3. Изменение темпа речи и четкости произношения является причиной коартикуляционной нестационарности, означающей изменение взаимовлияния соседних звуков от образца к образцу.

4. Проблема кластеризации слитной речи: в непрерывном речевом потоке трудно распознать речевые единицы из-за неточного определения границ.

Вот лишь некоторые причины, препятствующие полной реализации систем распознавания речи.

4. Области применения

Обозначим основные области применения систем распознавания речи:

1. *Автоматизированный пользовательский интерфейс.* На сегодняшний день для многих людей общение с компьютером все еще вызывает затруднения. Системы распознавания речи позволяют преодолевать эти трудности. Огромное преимущество систем распознавания голоса в том, что они намного быстрее любых других типов интерфейсов. Голосовая программа электронной почты позволяет включать компьютер, диктовать и послать сообщения, не прикасаясь к мыши и клавиатуре. Также люди с физическими недостатками получат более эффективный способ взаимодействия с компьютером.

Наиболее очевидное использование системы распознавания слитной речи заключается в создании систем автоматического стенографирования, которые могут заменять секретарей при диктовке голосом текстов писем, заметок в ежедневник, докладов. В таком случае происходит не только экономия за счет сокращения работы стенографиста, но и повышение степени конфиденциальности информации.

2. *Управление мобильными устройствами.* Известно, насколько неудобно и опасно использование мобильных телефонов с обычным (тактильным) способом набора номера за рулем. Во многих странах приняты законы о запрете использования водителями таких телефонов с целью сокращения количеств ДТП. Поэтому в последнее время популярностью пользуются мобильные телефоны с голосовым набором, избавляющие пользователя от необходимости набирать нужный номер вручную. Достаточно произнести имя абонента, и соединение произойдет автоматически. Аудиосистемы контроля и управления уже применяются в автомобилях некоторых производителей. Владелец автомобиля голосом подает команды управления температурным режимом, радио, навигационной системой, которые воспринимают голос и выполняют команды (DIVO и VoiceCommander).

3. *Информационные услуги.* Современные системы распознавания речи применяются, например, для заказа авиабилетов, просмотра новостей, доступа к базам данных.

Технология распознавания голоса быстро изменила рынок телефонных услуг. Системы, распознающие разговорную речь, работают в информационных телефонных центрах (IVR-системы – Interactive Voice Response). Эти системы позволяют автоматизировать диалог с клиентом, в результате чего отпадает необходимость в огромном количестве операторов, принимающих телефонные звонки, и сокращаются расходы на содержание персонала. Вдобавок улучшается качество обслуживания клиентов, так как соединение с машиной осуществляется практически сразу, избавляя клиентов от длительного ожидания освободившегося оператора на линии.

4. *Бизнес и профессиональная поддержка.* Уже многие годы голосовые диктофонные системы, предназначенные для представителей определенных профессий, например, врачей и юристов, можно найти на рынке программных продуктов. Многие представители этих профессий используют системы распознавания речи в повседневной работе. Стали популярны активируемые голосом домашние приборы и приспособления.

5. *Комбинированные человеко-машинные интерфейсы.* За последнее десятилетие области применения таких систем значительно расширились и будут продолжать расширяться. Они применяются, в частности, для контроля ограниченного доступа к объекту с помощью распознавания лица и речи человека, выполнения финансовых операций при помощи речи и сенсорных экранов банкоматов. В качестве примеров можно привести российско-белорусский проект “Модель аудиовизуального синтеза и распознавания речи для интеллектуальных устройств массового обслуживания” (2008-2009 гг), российско-турецкий проект “Методы и многомодальные интерфейсы для бесконтактной коммуникации инвалидов с информационно-справочными системами” (2009-2010 гг.), проект Российской академии наук “Разработка средств универсального многомодального доступа для системы интерактивного телевидения” (2009-2010 гг.).

5. Перспективы развития

Основными препятствиями на пути дальнейшего развития автоматизированных систем распознавания речи являются:

1) необходимость больших объемов словарей;

2) зашумленность речевого сигнала;

3) различные акценты и произношения.

Объемы словарей определяют степень сложности, требования к вычислительной мощности и надежность систем распознавания речи. Необходимо продолжать основательные исследования. Это

позволит решить проблемы, связанные с морфологией, акцентами, высотой звука, темпом, громкостью, сливающимися словами, артикуляцией, лингвистической информацией и т. д. Ожидается, что основным направлением развития станет моделирование языков для использования в системах распознавания речи.

Не решена окончательно и проблема выделения речевого сигнала из шумового фона. В настоящее время пользователи систем распознавания голоса вынуждены работать в условиях минимального шумового фона.

Одна из приоритетных разработок в области распознавания речи – это человеко-машинные диалоговые системы, работа над которыми ведется во многих исследовательских лабораториях мира. Одной из таких разработок является техническая система фирмы AT&T (США), которая используется для распознавания речи в телефонной сети: 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.

Такие системы "умеют" работать с непрерывным речевым потоком и с неизвестными дикторами, понимать значения фрагментов речи ограниченного словаря и предпринимать ответные действия. Системы работают в реальном времени и способны выполнять пять функций:

1) узнавание речи – преобразование речи в текст, состоящий из отдельных слов;

2) понимание – грамматический разбор предложений и распознавание смыслового значения;

3) восстановление информации – получение данных из оперативных источников на основании полученного смыслового значения;

4) генерация лингвистической информации – построение предложений, представляющих полученные данные, на выбранном пользователем языке;

5) синтез речи – преобразование предложений в синтезированную компьютером речь.

Диалоговый интерфейс в таких системах позволяет человеку разговаривать с машиной, создавать и получать информацию, решать свои задачи. Системы с диалоговым интерфейсом различаются по уровню инициативности человека или компьютера. Исследования фокусировались на "смешанно инициативных" системах, в которых как человек, так и компьютер играют одинаково активную роль в достижении цели посредством диалога.

Как ожидают в Datamonitor, одном из лидирующих мировых маркетинговых агентств, объем мирового рынка систем автоматического распознавания речи вырастет с $32,7 млн в 2009 г до $99,6 млн в 2014 г. Примерно теми же темпами будет расти и рынок систем распознавания для автомобильных телематических систем: с $64,3 млн в 2009 г. до $208,2 млн в 2014 г. "Рост популярности голосового интерфейса в телефонах будет расти по мере того, как все большее число их владельцев сталкиваются с необходимостью использовать мобильник в ситуациях, когда руки и глаза заняты", – говорят специалисты.

Заключение

Ограничения применения систем распознавания речи в рамках наиболее традиционных приложений позволяют сделать вывод о необходимости поиска потенциально новых решений в области распознавания речи. В ближайшее десятилетие задача распознавания и понимания естественной речи вне зависимости от языка и диктора будет занимать центральное место в речевых технологиях.

В настоящее время в ХНУРЭ разрабатывается новый метод автоматического распознавания речевых сигналов в реальном масштабе времени, основанный на бионическом принципе анализа сигналов.

Список литературы. 1. *Мясников Л.Л.* Звуки речи и их объективное распознавание // Вестник ЛГУ. 1946. – №3. 2. *Dudley H., Riesz R., Watkins S.* "A Synthetic Speaker" // Journal of the Franklin Institute. 1939, 227. — P. 739–764. 3. *Davies, K.H., Biddulph, R. and Balashek, S.* (1952) Automatic Speech Recognition of Spoken Digits, J. Acoust. Soc. Am. 24(6). — P. 637 – 642. 4. Вокодерная телефония. Методы и проблемы. /Под ред. А.А. Пирогова. – М: Связь, 1974. 5. *Клэтт Д.Х.* Основные результаты работ по проекту ARPA //Методы автоматического распознавания речи. М. — 1983. — Т. 1. 6. *Рабинер Л.* Скрытые марковские модели и их применение в избранных приложениях при распознавании речи: Обзор. ТИИЭР. — 1989, т. 77, №2. — С. 86-120. 7. *Винцюк Т. К.* Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 262 с. 8. *Секунов Н.* Обработка звука на PC. – СПб.: БХВ-Петербург. – С. 2001-1248.

Поступила в редколлегию 29.04.2010

УДК 004.934

Розпізнавання мови: етапи розвитку, сучасні технології і перспективи їх застосування / М.Ф. Бондаренко, А.В. Работягов, С.В. Щепковський // Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 2 (73). – С. 164–168.

Проводиться короткий огляд розвитку систем розпізнавання мови, розглянуті загальні принципи їх побудови, а також перераховані основні етапи розвитку цього напряму і актуальні проблеми, пов'язані з вирішенням завдань розпізнавання мови.

Бібліогр.: 8 найм.

UDC 004.934

Speech recognition: stages of development, modern technologies and prospects of their application / M. Bondarenko, A. Robotyagov, S. Schepkovsky // Bionics of Intelligence: Sci. Mag. – 2010. – № 2 (73). – C. 164–168.

The brief review of development of the systems of speech recognition is conducted, general principles of their construction are considered, and also the basic stages of development of this direction and issues of the day, related to the decision of tasks of speech recognition are transferred.

Ref.: 8 items.