УДК:004.627

Технологии распознавания речи

Клевцова А.С.

Аннотация : После появления аппаратных средств по обработке звука с помощью компьютера стали возможными попытки создания интерфейса, основанного на речевых технологиях – *речевого интерфейса* или *SILK-интерфейса* (*Speech, Image, Language, Knowledge* – речь, образ, язык, знание). Речевой интерфейс нужен для повышения удобства и интеллектуализации человеко-машинного диалога. Речевое управление удобно и полезно в тех случаях, когда руки и зрение пользователя заняты важными «неинтерфейсными» задачами: вождение транспорта, визуальный осмотр, просмотр фильма, тонкие манипуляции и т.п.

Важной задачей при создании речевого интерфейса является задача распознавания речи, которая будет рассмотрена в настоящей статье.

Ключевые слова: речевой интерфейс, дискретизация, голосовой помощник, распознавание речи, синтез речи.

Распознавание речи – это многоуровневая задача распознавания образов, акустические сигналы которой, анализируются и структурируются в иерархию элементов (например, фонем), слов, фраз и предложений.

Что касается распознавания речи, то в рассматриваемой задаче она сводится к процессу извлечения из речи текста.

Традиционно процесс распознавания речи подразделяется на несколько этапов, соответствующих блокам системы распознавания речи:

1. *Дискретизация необработанного речевого сигнала, преобразованного в электрическую форму*. Обычно частота дискретизации составляет 20 кГц при записи с микрофона, 8 кГц при записи с телефонной линии.

2. *Анализ сигнала*. Дискретный речевой сигнал подвергается очистке от шумов и преобразуется в более компактную форму (сжимается). Сжатие производится посредством вычисления каждые 10 мс некоторого набора

числовых параметров с минимальными потерями информации, описывающей данный речевой сигнал. К наиболее часто используемым для этого методам относятся: *линейно-предиктивное кодирование*, *преобразование Фурье*, *кэпстральный анализ*. Результатом анализа сигнала является последовательность речевых кадров, содержащих информацию, необходимую для последующего распознавания.

3. *Распознавание сигнала*. Хранимые в памяти компьютера эталонные произношения (акустические модели) сопоставляются с каждым речевым кадром, и формируется матрица сопоставления последовательности кадров и множества акустических моделей.

Наиболее часто используются два вида акустических моделей:

- *Шаблонная модель*. Сохраняется несколько вариантов произношения одного и того же элемента (множество дикторов повторяют одну и туже команду). Используется, в основном, для распознавания слов, как единого целого.

- *Модель состояний*. Каждое слово моделируется как последовательность состояний, указывающих набор звуков, которые возможно услышать в данном участке слова, основываясь на вероятностных правилах.

Для шаблонной модели матрица сопоставления представляет собой Евклидово расстояние (минимальное) между шаблонным и распознаваемым кадром. Для модели состояний матрица состоит из вероятностей того, что данное состояние может сгенерировать данный кадр.

В зависимости от степени совпадения выбирается лучший вариант и формируется гипотеза о содержании высказывания. Здесь может возникнуть проблема – необходимость нормализации сигнала по времени, поскольку темп речи, длительность произношения отдельных слов и звуков даже для одного диктора варьируется в широких пределах, что может привести к значительным расхождениям между отдельными участками акустической модели и теоретически совпадающим с ней входным сигналом. Достаточно эффективно данная проблема решается с помощью алгоритма динамического

программирования (алгоритма Витерби) путём сжатия и растяжения сигнала по временной оси непосредственно в процессе сравнения или Марковских моделей, позволяющие производить временную нормализацию и прогнозирование продолжений, что ускоряет процесс перебора эталонов и повышает надёжность распознавания.

На рынке представлено множество коммерческих систем распознавания речи, в частности, VoiceTypeDictation, VoicePilot от IBM, VoiceAssistCreative от Technology, NUANCE Dragon, программные продукты от ООО «Центр речевых технологий», облачные сервисы GoogleSpeech, Yandex.SpeechKit, Siri, MSCortana и многие другие.

Технологии и системы распознавания речи применяются сейчас в различных сферах человеческой деятельности. Многие пользователи решают свои задачи с помощью мобильного телефона: просматривают почту, отправляют документы, фотографии, осуществляют поиск информации и т.п., для чего удобно использовать голосовое управление.

Ниже рассмотрим некоторые программные продукты для мобильных устройств, основанных на технологии распознавания речи.

На конференции YetAnotherConference в октябре 2013 года компания Yandex представила свою технологию распознавания речи *SpeechKit*, которая может использоваться на мобильных устройствах под управлением Android, iOS, WindowsPhone и позволяет взаимодействовать с мобильными приложениями Yandex. В ней используется вероятностная акустическая модель на основе нейронных сетей, для выравнивания сигнала во времени – алгоритм Витерби. По данным компании, *SpeechKit* позволяет правильно распознавать 94% слов в Навигаторе и мобильных картах, а также 84 % слов в Мобильном Браузере. При этом на распознавание уходит чуть больше секунды.

На основе этой технологии Yandex был разработан облачный сервис распознавания речи *SpeechKit Cloud*, который она представила в 2014 году. С его помощью разработчики могут научить свои продукты понимать голос человека. В поддержку *SpeechKit Cloud* можно добавить в самые разные

программы, сервисы и устройства: от компьютерной игры до автомобильной навигационной системы.

SpeechKit Cloud понимает русский и турецкий языки. Обработка голосовых запросов производится на серверах Yandex, рассчитанных на высокие нагрузки. Распознавание занимает около секунды: именно столько времени проходит с момента отправки данных на сервер до получения ответа.

Взаимодействие со *SpeechKitCloud* ведётся через HTTPAPI. Это позволяет значительно сократить время интеграции и применять технологию распознавания речи Yandex в различных сценариях: *для ПК и ноутбуков, для автомобилей, в телефонии, в медицине, дома, для бытовой техники* и др.

У Yandex есть несколько партнёров, которые уже используют *SpeechKit Cloud* в своих продуктах. Один из партнёров - это компания Cubic Robotics. Она разрабатывает домашнего робота-помощника CUBIC. Используя технологию Yandex, CUBIC распознает вопросы человека и отвечает на них. Он может, к примеру, включить или выключить свет в комнате, зачитать последние новости или рассказать о пробках на дорогах.

Программа *SpeechKit Mobile SDK* позволяет встроить распознавание и синтез речи, а также голосовую активацию Яндекса в мобильное приложение для iOS, Android или Windows Phone.

Используется: в *Яндекс.Навигатор*, *Яндекс.Браузер*, в приложении *Aviasales*.

SpeechKit Box позволяет реализовывать функции распознавания и синтеза речи, а также смыслового разбора сказанного в сервисах и приложениях. Комплекс разворачивается во внутренней сети клиента, так что данные не передаются для обработки на внешние сервера. Благодаря этому речевые технологии можно использовать для работы с конфиденциальной информацией.

Приложение *Яндекс. Диктовка* было представлено на конференции Yet Another Conference в Москве. Технология позволяет запускать голосовой ввод с помощью активационной команды «*Яндекс, записывай*». Сообщается, что с

помощью сервиса пользователи смогут набирать, например, СМС, письма или комментарии в социальных сетях.

Яндекс.Диктовка работает с использованием трех новых функций технологии распознавания речи, разработанной компанией. Эти функции обеспечивают голосовую активацию работы приложения, отвечают за пунктуацию и синтез.

По окончанию диктовки, пользователь может заслушать результат расшифровки и, в случае необходимости, отредактировать его с помощью специальных голосовых команд. Например, *«Удали последнюю фразу»* или *«Сотри последнее предложение»*.

4 октября 2011 Apple проводит презентацию iPhone 4S, где представляет бета-версию интеллектуального голосового помощника *Siri (Speech Interpretation and Recognition Interface)*, которая стала не только ключевым нововведением iOS5, но и эксклюзивом для iPhone 4S.

Siri является результатом работы исследований, которые накапливались более сорока лет. В 2007 году начала работу над *Siri* компания SRI International, являющаяся подразделением *DARPA* (Агентство по перспективным оборонным научно-исследовательским разработкам). Позже Siri купила компания Apple.

Siri интегрирована в iPhone 4S, взаимодействует с основными приложениями iOS, отвечает на трудные вопросы.

В основе *Siri* лежат технологии, способные понимать естественный язык, машинное обучение, технологии очевидного и вероятного рассуждения, онтологии, представления знаний и планирования. *Siri*, по сути, является усовершенствованной версией *CALO (Cognitive Agent that Learns and Organizes Program)*, познавательного агента, программы, способной к самообучению и организации.

Siri ведет полноценный диалоговый разговор с пользователем. Кроме этого, Siri использует совершенствующуюся технологию распознавания речи человека, которую разработала компания Nuance Communications. *Siri*

индивидуально приспосабливается к каждому пользователю: слушает и изучает своего владельца, анализируя его предпочтения.

В отличие от других голосовых помощников, работающих просто с поисковой системой, *Siri* работает с множеством сервисов, что позволяет точно отвечать на самые разные вопросы, в том числе и очень сложные. После формулировки запрос отравляется на сервера Apple (Siri), где обрабатывается и направляется к соответствующему сервису. Помимо универсальных поисковых систем, например, Google, Bing, *Siri* использует специализированные сервисы. Например, для деловых вопросов используются OpenTable, Andre Gayot, Citysearch, BooRah, Yelp Inc, Yahoo Local, ReserveTravel и Localeze. Для поиска информации о мероприятиях *Siri* обращается к Eventful, StubHub и LiveKick. Если спрашивать Siri о фильмах, то она отвечает, используя информацию с MovieTickets.com, Rotten Tomatoes и The New York Times.

Таким образом, голосовой ассистент от Apple справится с большинством повседневных вопросов, но ключевой особенностью является то, что *Siri* работает с WolframAlpha. WolframAlpha позволяет *Siri* давать ответы на самые трудные вопросы, поскольку она позиционирует себя, как *computational knowledge engine* (база знаний и набор вычислительных алгоритмов).

Благодаря всему вышеперечисленному, *Siri* удается понимать речь человека и его вопросы, которые он задает в достаточно свободной форме, а не конкретные команды. Изданием PhoneArena были протестированы голосовые помощники *Siri*, *GoogleNow* и *MSCortana*. В результате тестирования, в ходе которого были заданы вопросы на пяти языках, выяснилось, что средняя доля распознаваний на всех языках составила 76% у *Siri* против 46% у *Cortana* и 42% у *GoogleNow*.

В апреле 2014 года компания Microsoft представила голосового помощника *Cortana* для мобильных устройств на операционной системе WindowsPhone.

Cortana начали разрабатывать в июне 2013 года. В компании сообщили, что приложение планируется внедрить в мобильную и настольную версию

Windows и в Xbox One, а также в приложении Bing для iOS и для Windows 10; приложение будет «говорить» голосом Джен Тейлор (Jen Taylor), которая озвучивала героя Cortana в играх Halo, отсюда и пошло название помощника.

Среди остальных возможностей и функций *Microsoft Cortana* – стандартный прием и обработка команд и вопросов, которые произносят пользователи. Ассистент сможет отвечать в голосовом виде или в письменном. Также данные социальной сети Foursquare будут использоваться в службах геолокации мобильной версии Windows. Русская локализация отсутствует полностью.

Первыми устройствами, на которых появилась предустановленная Cortana, стали смартфоны линейки Nokia Lumia: они получили приложение в статусе бета уже в апреле 2014 года. Генеральный директор Стив Балмер в то время заявил, что в данный момент у Cortana улучшается распознавание речи, а разрабатываемые технологии Cortana, которые основываются на поисковике Bing, станут главными элементами сервиса. Также он дополнил заявление тем, что приложение будет персональным, иметь интеллект и уметь собирать информацию со всех сервисов Microsoft, объединяя ее для раскрытия потенциала ассистента.

В январе 2016 года голосовой помощник *Cortana* установлен на некоторых устройствах OnePlusOne.

Таким образом, несмотря на существующие проблемы (невозможность подавления внешних шумов, невысокая точность распознавания голоса, высокая стоимость программных приложений распознавания голоса), технологии распознавания речи развиваются достаточно быстрыми темпами и внедряются в различные сферы человеческой деятельности, в частности, здравоохранение, финансы, военную сферу. Для широкого круга потребителей наиболее перспективным направлением развития систем распознавания голоса является их применение в мобильных устройствах, о чём свидетельствуют рассмотренные в статье отечественные и зарубежные технологии распознавания речи.

Список литературы

1. Using Cortana and Speech Recognition Together on Windows 10 [Электронныйресурс]. – Режимдоступа: http://www.equalentry.com/using-cortana-and-speech-recognition-together-on-windows-10/. – Систем. требования: P IV; 64 МбОЗУ; Windows 98 ивыше; SVGA 32768 иболеецветов; 640×480; мышь; IE 4.0 ивыше. – Загл. сэкрана.

2. Siri/ VoiceRecognitionSystem [Электронный ресурс]. – Режим доступа: http://techinfo.subaru.com/proxy/105404/pdf/ownerManual/105404_2016_Legacy/MSA5M1611ASTIS052215_12.pdf. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

3. Голосовой помощник MicrosoftCortana [Электронный ресурс]. – Режим доступа: http://wd-x.ru/znakomtes-golosovoj-pomoshhnik-microsoft-cortana. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

4. Перспективы речевого интерфейса [Электронный ресурс]. – Режим доступа: http://www.f-mx.ru/kompyutery_evm/rechevye_texnologii.html. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

5. Перспективы развития систем распознавания речи [Электронный ресурс]. – Режим доступа: http://savepearlharbor.com/?p=232613. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

6. Распознавание речи от Яндекса [Электронный ресурс]. – Режим доступа: http://4pda.ru/2014/08/04/169812/. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

7. Речевые технологии SpeechKit Яндекса [Электронный ресурс]. – Режим доступа: tech.yandex.ru›РечевыетехнологииSpeechKit. – Систем.

требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

8. Состояние исследований в области распознавания речи [Электронный ресурс]. – Режим доступа: http://www.auditech.ru/page/galunov_d.html. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

9. Типичная структура системы распознавания речи [Электронный ресурс]. – Режим доступа: http://uc.org.ru/node/60. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.

10. Что такое Siri на iPhone и как она работает? [Электронный ресурс]. – Режим доступа: http://app-s.ru/publ/chto_takoe_siri_i_kak_ona_rabotaet/1-1-0-33. – Систем. требования: P IV; 64 Мб ОЗУ; Windows 98 и выше; SVGA 32768 и более цветов; 640×480; мышь; IE 4.0 и выше. – Загл. с экрана.