

ОБЗОР СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

Локтионов Никита Павлович

магистрант, ИСОиП (ф) ДГТУ, РФ, г. Шахты

Ляшов Максим Васильевич

научный руководитель, канд. техн. наук, доц., ИСОиП (ф) ДГТУ, РФ, г. Шахты

Стремительное развитие голосового поиска и систем распознавания речи играет важную роль в создании и усовершенствовании современных поисковых систем и систем управления. Управление с помощью голоса постепенно выходит за пределы планшетов и смартфонов и становится важной частью повседневной жизни. В перспективах развития – распознавание акцентов и диалектов, абсолютно новые подходы к формированию и выдаче информации, создание приложений, полностью управляемых голосом. Необходимо также отметить, что акценты, сленг и речевые модели достаточно разнообразны и на сегодняшний день все еще сложны для систем распознавания речи, но точность распознавания из года в год становится лучше. Осенью 2015 года Google представил улучшенные акустические модели, созданные с помощью технологии CTC (Connectionist Temporal Classification, нейросетевая темпоральная классификация). Это позволило повысить точность в условиях посторонних шумов, а также увеличить скорость распознавания.

Разработка системы распознавания речи – очень ресурсозатратная и довольно трудоемкая задача. Поэтому логично предположить, что для собственного проекта нужно интегрировать в него готовый вариант системы. В общем случае системы распознавания речи можно разделить на две категории: с открытым исходным кодом и с закрытым исходным кодом. У каждой категории есть свои достоинства и недостатки. Так, реализация систем распознавания речи с закрытым исходным кодом выполнена более качественно, такие системы имеют высокую точность распознавания речи, однако они могут не иметь необходимой документации, связанной с интегрированием других решений в их работу. Часто подобные системы являются платными, то есть придется покупать лицензию, чтобы использовать речевые технологии, представленные компанией-разработчиком. Системы распознавания речи с открытым исходным кодом обычно бесплатны, но для их эффективной работы необходимо создание исходной библиотеки с расширенным набором артикуляционных данных, речевых единиц и т.п. Это означает, что за развитие таких открытых систем ответственны те, кто их использует.

Рассмотрим системы распознавания речи с закрытым исходным кодом и возможностью интеграции. Мировым лидером в создании ПО для распознавания речи является Nuance Communications. Инструментарий Dragon Mobile SDK состоит из компонентов клиента и сервера, а также содержит в себе разнообразные примеры кода и шаблоны, документацию и фреймворк, значительно упрощающий интеграцию сервисов в приложения. Платформа Speech Kit создана для простого и быстрого добавления сервисов распознавания и синтеза речи в проекты и приложения. Также с помощью данной платформы можно получить доступ к серверным компонентам обработки речи. Все это минимизирует потребляемые ресурсы и расходы.

За большую часть операций цикла обработки речи отвечает система серверов. На сервере полностью выполняется распознавание или синтез речи. С помощью сервера выполняется аутентификация, зависящая от конфигурации разработчика.

Платформа Speech Kit – это, по сути, сетевой сервис, нуждающийся в базовых настройках для возможности использования классов распознавания или синтеза. Для этого выполняются основные операции:

- Определение и авторизация приложения;
- Установка соединения с сервером обработки речи.

Все это обеспечивает создание быстрых запросов на обработку речевой информации и повышает качество работы.

Dragon Mobile SDK показывает высокую точность распознавания на английском языке (до 99%). Его недостаток – ограниченный бесплатный функционал (до десяти тысяч запросов в сутки, выше десяти тысяч – платный доступ).

Google Speech Recognition API – продукт компании Google, позволяющий использовать голосовой поиск на основе технологии распознавания речи. Эта технология интегрирована в смартфоны и компьютеры с возможностью ввода голосовой информации. Летом 2011 года корпорация интегрировала речевую технологию в свою поисковую систему. На персональных компьютерах поддерживается браузером Google Chrome. Также существует голосовое управление для смартфонов с ОС Android.

С мая 2014 года доступ к API стал легален. Для работы с базой данных системы распознавания необходимо зарегистрироваться в Google Developers. Голосовой поиск Google по умолчанию внедрен во многие популярные сервисы: Google, Yahoo, YouTube, Википедия и др. Имеется возможность добавления собственных поисковых систем. Существует расширение, которое добавляет кнопку ввода голосовой информации для сайтов, использующих поисковые формы HTML5.

Чтобы использовать технологию голосового поиска Google необходимо выполнить POST-запрос на адрес со звуковыми данными формата «.flac» или «.spx». Далее реализовать распознавание WAVE-файлов с помощью любого ЯП. Система распознавания от Google во многом похожа на Dragon Mobile SDK от Nuance, но при этом не имеет ограничений на количество запросов в сутки.

Yandex Speech Kit – разработчики уверяют, что этот SDK является наилучшим выбором при использовании русского языка. Установлено ограничение на количество запросов (десять тысяч в сутки). По словам Григория Бакунова (топ-менеджер Яндекса), такое решение «достаточно тяжело вывести на международный рынок, так как с точки зрения патентов многое принадлежит Nuance и Яндекс одним из последних сумел продвигать развитие систем распознавания речи».

Эффективность распознавания в Yandex Speech Kit зависит, в первую очередь, от качества исходного звука, кодирования, разборчивости речи, темпа речи, сложности и длины фраз. Тематика голосовых запросов должна максимально совпадать с выбранной языковой моделью. Распознавание речи происходит в реальном времени вместе с передачей звуковой информации. Задержка не превышает одной секунды. Для обеспечения такой высокой скорости технология работает в режиме потокового распознавания с промежуточными результатами. Это значит, что как только человек начинает говорить, его речь передается на сервис распознавания небольшими частями.

Microsoft Speech API – технология распознавания речи компании Microsoft. В последнее время корпорация стала активно развивать подобные технологии (анонс Cortana, разработка технологии синхронного перевода в Skype).

Существуют различные варианты использования API:

- Windows и Windows Server – добавление речевой технологии для Windows-приложения с использованием управляемого или исходного кода, который берется из API, а также управление речевым движком, встроенным в Windows и Windows Server;

- Speech Platforms – встраивание платформы в различные приложения, использующие дистрибутивы Microsoft (например, встраивание в языковые пакеты);
- Embedded – решения, позволяющие человеку взаимодействовать с устройствами, используя голос (например, управление в ОС Windows Automotive);
- Services – разработка приложений голосовыми функциями для использования в реальном времени.

Выше были рассмотрены одни из самых распространенных систем распознавания речи с закрытым исходным кодом. Какую же систему наиболее выгодно использовать в своем проекте? Привлекательно выглядит инструментарий Dragon Mobile SDK (хорошая документация, простой код для встраивания). Однако данный продукт имеет сложную систему лицензирования и строгие правила использования технологии. Поэтому целесообразность создания пользовательского продукта на Dragon Mobile под вопросом.

Исходя из вышесказанного, наиболее выгодно использование Google Speech Recognition API (хорошая встраиваемость и быстроедействие за счет больших вычислительных мощностей корпорации). Еще одно преимущество – отсутствие ограничения по количеству запросов в сутки. Google довольно активно развивает технологии распознавания речи, что также является плюсом при использовании их инструментария.

Теперь рассмотрим системы распознавания речи с открытым исходным кодом. CMU Sphinx (или просто Sphinx) была создана группой разработчиков из университета Карнеги-Меллон. Система состоит из набора распознавателей речи (Sphinx 2-4) и акустической модели (Sphinx train).

Sphinx является распознавателем непрерывной речи, использующим скрытые марковские модели и статистические языковые модели. В системе реализована возможность распознавания продолжительной речи и большой словарь распознавания. Sphinx примечателен тем, что развиваясь, затмевал все предыдущие версии в плане производительности.

Sphinx 4 – абсолютно новый речевой движок Sphinx, обеспечивающий основу для исследования технологий распознавания речи. Sphinx 4 был создан на ЯП Java. Большой вклад в создание и развитие Sphinx 4 внесла и сама Sun Microsystems. В будущем разработчики планируют создание и развитие новых акустических моделей, реализацию системы речевой адаптации, улучшение управления конфигурациями.

PocketSphinx может быть встроена в системы на базе ARM-процессора. На данный момент PocketSphinx быстро развивается и встраивается в различные системы и модели на базе смешанных вычислений.

Julius – высокопроизводительная система распознавания речи с большим словарем, также декодер ПО для исследования связанной речи. Декодер работает на большинстве современных компьютеров, объем словаря – 60 тысяч слов. Главная особенность системы – полная встраиваемость. Основной платформой проекта являются системы Linux и другие UNIX-подобные системы, однако существует версия и для Windows.

Изначально Julius разрабатывался для исследований в области распознавания японского языка, разработки продолжались в рамках консорциума систем распознавания непрерывной речи с 2000 по 2003 год в Японии. С версии 3.4 грамматическая база анализатора получило название Julian и была интегрирована в Julius. Julian использует собственную форму грамматики на основе конечного автомата и является, по сути, модифицированной формой Julius. Существенный недостаток Julius – распространение только для модели японского языка. Стоит заметить, что существует проект VoxForge, который создает акустическую модель для английского языка на основе Julius.

RWTH ASR (RASR) – система распознавания речи, включающая в себя технологию для создания автоматических систем распознавания. Эта технология развивается группой из

Рейнско-Вестфальского технического университета Ахена. RWTH ASR состоит из инструментов для создания акустических моделей и декодеров, а также компонентов для адаптации речи и адаптивных систем обучения. Данная система распознавания речи работает с операционными системами Linux и MacOS. Проект имеет подробную документацию и предлагает готовые для использования шаблоны и модели, обучающие системы и др.

Simon – система распознавания речи, созданная с помощью речевых движков Julius и HTK. Simon довольно удобна для работы с различными языками и диалектами. Работа системы основана на выполнении определенных сценариев. Пакеты системы сконфигурированы для выполнения специализированных задач. Пользователи могут сами создавать сценарии и выкладывать их в сообщество. Simon поддерживает разнообразные модели, подобные GPL-моделям из VoxForge. Пользователи используют эти модели для произношения. Сама система может работать без предварительной тренировки.

Рассмотрев распространенные системы распознавания речи с открытым исходным кодом можно сделать вывод, что наиболее интересна система CMU Sphinx. Однако, как было сказано ранее, для эффективной работы систем распознавания речи с открытым исходным кодом необходима большая база исходных данных, в противном случае точность распознавания будет небольшой, особенно в сравнении с системами с закрытым исходным кодом. В общем случае использование подобных систем как основы для собственной системы распознавания речи вполне оправдано.

Список литературы:

1. Голосовой поиск в Google Chrome – [Электронный ресурс] // Habrahabr. – Режим доступа. – URL: <http://habrahabr.ru/post/111201> (дата обращения 10.02.2017).
2. Kai Fu Li, Hsiao-Wuen Hon. An overview of the Sphinx Speech Recognition Systems – [Электронный ресурс]. // The Robotics Institute – Carnegie Mellon University. – Режим доступа. – URL: http://www.ri.cmu.edu/pub_files/pub2/lee_k_f_1990_1/lee_k_f_1990_1.pdf (дата обращения 10.02.2017).
3. Nuance Developers – [Электронный ресурс]. // Nuance Communications, Inc. – Режим доступа. – URL: <https://developer.nuance.com/public/index.php?task=home> (дата обращения 10.02.2017).
4. Speech Input API Specification – [Электронный ресурс]. // W3C. – Режим доступа. – URL: <https://www.w3.org/2005/Incubator/htmlspeech/2010/10/google-api-draft.html> (дата обращения 10.02.2017).