электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.934.1'1

Исследование сервиса Google Voice как средство распознавания речи с применением программных технологий

Груздев О.В., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Информационные системы и телекоммуникации»

Научный руководитель: Селихов Ю.Р. доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
deviatkov@bmstu.ru

Распознавание речи - технология, позволяющая использовать естественный для человека речевой интерфейс при взаимодействия с электронной техникой. По мере развития IT технологий становится все более очевидным, что использование систем голосового распознавания сильно расширится, если станет возможным использование человеческой речи при работе с компьютером, и в частности станет возможным управление компьютером обычным голосом в реальном времени, ввод и вывод информации в виде человеческой речи. Сегодня системы распознавания речи основываются на сборе всей доступной информации, необходимой для распознавания слов. Исследователи считают, что таким образом задача распознавания образца речи, основанная на качестве сигнала, подверженного изменениям, будет достаточной для распознавания, но тем не менее в настоящее время даже при распознавании небольших сообщений нормальной речи, пока невозможно после получения разнообразных реальных сигналов осуществить прямую трансформацию в лингвистические символы, что является желаемым результатом. Вместо этого проводится процесс, первым шагом которого является первоначальное трансформирование вводимой информации для сокращения обрабатываемого объема так, чтобы ее можно было бы подвергнуть компьютерному анализу.

В связи с активным внедрением электронной техники в жизни людей, появилась необходимость использования человеческой речи для облегчения управления электронными устройствами, бытовой техникой, что говорит об актуальности этой темы. В частности появились широко используемые веб-сервисы, позволяющие через определенный интерфейс (например Google Android Speech API для операционной системы Android или Google Voice) взаимодействовать с системой и осуществлять

преобразование речевой информации в текстовую для последующего использования в электронных устройствах и информационных системах. На сегодняшний день существует большое количество программного обеспечения позволяющего осуществить операцию преобразования голоса в текст, но зачастую оно занимает не мало памяти, не обладает гибкостью при использования его на различных операционных системах, а зачастую его использование просто невозможно из-за недостаточного количества памяти на современных устройствах (смартфоны, коммуникаторы, GPS навигаторы). Иной подход к обеспечению голосового распознавания дают интернет сервисы по распознаванию речи. Одним из наиболее мощных в настоящее время является сервис компании Google, встроенный в браузер Chrome интерфейс обеспечивает возможность записи речевой информации при наличии у клиента звукозаписывающего оборудованию и преобразование ее в текстовую, которая тут же используется по назначению так, как будто пользователь ввел ее сам, при помощи устройства ввода текстовой информации.

Для реализации голосового распознавания на данном этапе исследования, мы будем использовать инструменты Google Speech API. В настоящее время использование этой технологии доступно лишь при использовании браузера Google Chrome, а так же в операционной системе Android при распознавании поисковых голосовых запросов с использованием программного интерфейса Android Speech. Однако при детальном исследовании пути запроса из браузера, можно заметить, что после получения голосовых данных браузер направляет POST запрос, одному из серверов Google, данные для которого отправляются в формате FLAC.

FLAC (англ. Free Lossless Audio Codec) — популярный свободный кодек, предназначенный для сжатия аудиоданных без потерь. В отличие от аудио-кодеков, обеспечивающих сжатие с потерями (MP3, AAC, WMA, Ogg Vorbis) FLAC, как и любой другой lossless-кодек, не удаляет никакой информации из аудиопотока и подходит как для прослушивания музыки на высококачественной звуковоспроизводящей аппаратуре, так и для архивирования аудиоколлекции. На сегодняшний день формат FLAC поддерживается множеством аудиоприложений, а также имеет большое количество аппаратных реализаций.

Для того, что бы реализовать распознавание предварительно записанной голосовой информации нам необходимо сформировать POST запрос на один из серверов Google. Данное действие можно осуществить по разному, но т.к. нам необходимо гибко отслеживать параметры передачи запроса, получения ответа от сервера и при этом отследить время реакции сервиса мы воспользуемся библиотекой Apache HttpComponents Client, фирмы Apache. Данная библиотека способна предоставить гибкий доступ к

данным через http протокол, а так же управление запросами, ответами, формированием данных.

Технология работы тестовой программы проста и состоит из нескольких этапов:

1. Запись звукового файла в формате FLAC с определенным битрейтом.
2. Формирование POST запроса со звуковыми данными.
3. Выполнение запроса и считывание ответа.
4. Разбор ответа.

Ответ от сервера приходит в удобном формате JSON. JSON (англ. JavaScript Object Notation) — текстовый формат обмена данными, основанный на JavaScript и обычно используемый именно с этим языком. Как и многие другие текстовые форматы, JSON легко читается людьми. Несмотря на происхождение от JavaScript (точнее, от подмножества языка стандарта ECMA-262 1999 года), формат считается языконезависимым и может использоваться практически с любым языком программирования. Для многих языков существует готовый код для создания и обработки данных в формате JSON. За счёт своей лаконичности по сравнению с XML, формат JSON может быть более подходящим для сериализации сложных структур. Если говорить о веб-приложениях, в таком ключе он уместен в задачах обмена данными как между браузером и сервером (AJAX), так и между самими серверами (программные HTTP-интерфейсы). Формат JSON также хорошо подходит для хранения сложных динамических структур в реляционных базах данных или файловом кэше. Поскольку формат JSON является подмножеством синтаксиса языка JavaScript, то он может быть быстро десериализован встроенной функцией eval(). Кроме того, возможна вставка вполне работоспособных JavaScript-функций.

После исполнения кода, в консоль выводились соответствующие характеристики звукового файла и ответ от сервера в следующем представлении:

954-time

24-bitrate

answer:{"status":0,"id":"86df06e4ea4deec42dc7dc50e98251301","hypotheses":[{"utterance":"кухонный комбайн","confidence":0.91519046}]}, где значение ключа confidence : 0.91519046 означает вероятность того, что слово распознано правильно, utterance : кухонный комбайн - два слова, которые были записаны в формате FLAC с битрейтом 24 Кб/с, time - время в миллисекундах, за которое осуществляется запрос к серверу, обработка данных на сервере и получение ответа в клиентской части, реализованной нами.

Изначальный код был несущественно изменен, добавлены циклические запросы к серверу, для того, что бы определить среднее время ответа сервера для каждого значения битрейта с которым записан звуковой файл. Проведя статистические исследования получили средний результат от 1200 испытаний (таблицы 1 и 2) по 100 испытаний для каждого значения битрейта файла.

Таблица 1

Результаты экспериментальных исследований, опыты с 1 по 600

Битрейт звукового файла, Кб/сек	Время реакции сервера Google, Мс	Размер звукового файла, Кбайт	Усредненная скорость передачи, Кб/сек
80	1540	69.798	6182.015
56	1525	83.578	6306.955
48	1400	92.222	6569.168
40	1404	94.297	6052.524
32	1495	95.713	6114.187
24	1422	99.763	6133.414

Таблица 2

Результаты экспериментальных исследований, опыты с 601 по 1200

Битрейт звукового файла, Кб/сек	Время реакции сервера Google, Мс	Размер звукового файла, Кбайт	Усредненная скорость передачи, Кб/сек
80	1584	69.798	6180.012
56	1415	83.578	6206.555
48	1204	92.222	6519.462
40	1570	94.297	6122.163
32	1399	95.713	6014.587
24	1553	99.763	6031.316

Как видно из таблиц время распознавания системой звукового файла варьируется от 1.2 секунды. до 1.55 секунды. Из графиков (рисунок 1 и 2) можно заметить, что в обоих экспериментах среднее время ответа от сервера было минимальным при реакции на звуковой файл записанный с битрейтом 48 Кб/сек. Для того, что бы нагляднее изобразить зависимость скорости реакции Google на звуковые файлы с различным битрейтом, построим графики зависимости времени распознавания (ответа) от битрейта звуковых файлов (изображение на рисунке 1 и 2). Заметим, что каждая точка графика отражает усредненный результат из 100 испытаний проводившихся в рамках исследования скорости реакции Google на запрос распознавания речевой информации.

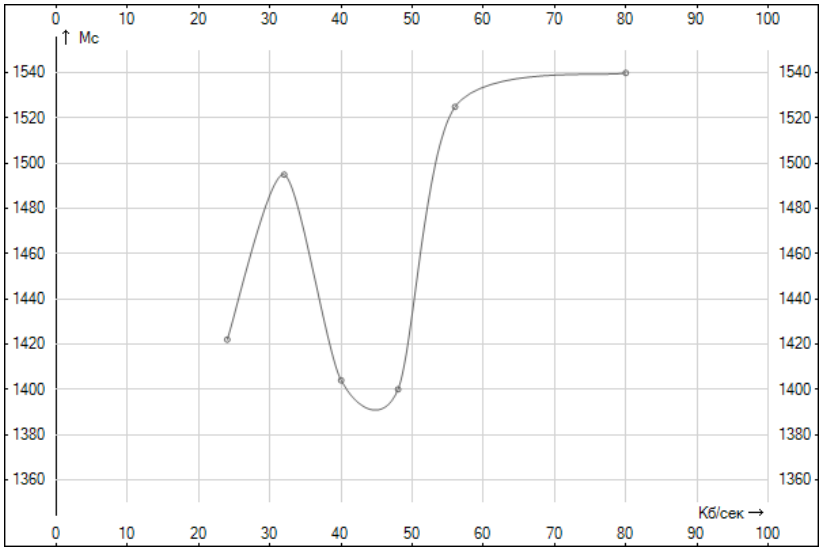

Рис. 1. График зависимости скорости реакции Google от битрейта звукового файла (к таблице 1)

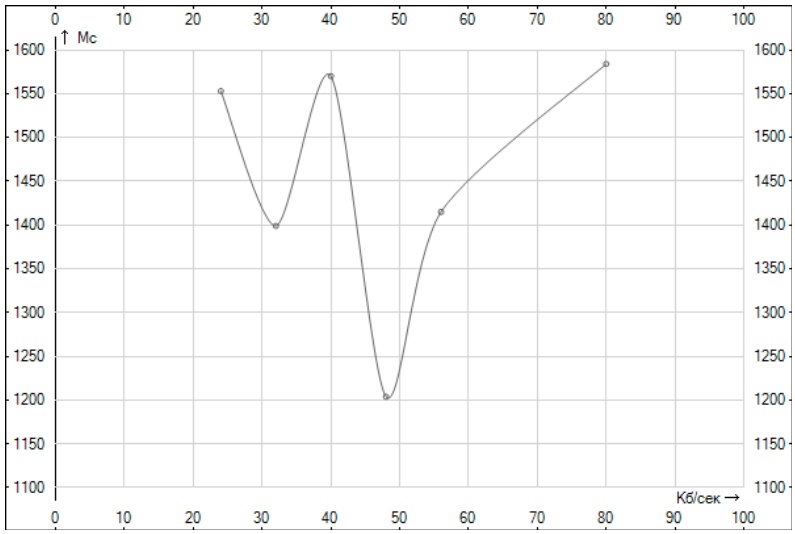

Рис. 2. График зависимости скорости реакции Google от битрейта звукового файла (к таблице 2)

В данной работе мы реализовали программно возможность использовать сервис голосового распознавания Google. Проведя исследования на предмет зависимости времени ответа сервера от битрейта звукового файла отправляемого серверу Google, мы установили, что в среднем из 1200 испытаний наиболее короткое время ответа получалось при отправке серверу звукового файла при записи с битрейтом 48 КБит/сек, однако, какие именно факторы влияют на время и качество распознавания звуковой информации сервисом Google, а так же их точное количество еще только предстоит выяснить.

Список литературы

1. Яхнев А.А. Использование Google Voice Search в приложении .NET. URL: http://habrahabr.ru/blogs/google/17234/
2. Материалы с сайта Википедия. [Электронный ресурс]. URL: http://ru.wikipedia.org/ (дата обращения: 13.05.2013г.).
3. А.В. Фролов, Г.В. Фролов, Синтез и распознавание речи. Современные решения, "Москва" 2008г.