REPÚBLICA BOLIVARIANA DE VENEZUELA
LA UNIVERSIDAD DEL ZULIA
FACULTAD EXPERIMENTAL DE CIENCIAS
DIVISIÓN DE PROGRAMAS ESPECIALES
LICENCIATURA EN COMPUTACIÓN

Arquitectura para la Edición de Mensajes de Texto en Dispositivos Móviles Utilizando Reconocimiento Automático de Voz

Proyecto de Trabajo Especial de Grado presentado como requisito para optar al título de Licenciado en Computación

Autor: Gerardo José Curiel Orosco

Tutor: MSc. Gerardo Pirela

Cotutor: MSc. Carlos Rincón

Maracaibo, junio de 2012

Arquitectura para la Edición de Mensajes de Texto en Dispositivos Móviles Utilizando Reconocimiento Automático de Voz

Gerardo José Curiel Orosco
CI. No.: 17.684.547
Teléfono: 58-261-7573725
Urb. La Paz
Correo electrónico: gcuriel@gmail.com

MSc. Gerardo Pirela
C.I. No.: 12.404.565
Correo electrónico: gepirela@fec.luz.edu.ve

MSc. Carlos Rincón
C.I. No.: 12.590.157
Correo electrónico: crincon@fec.luz.edu.ve

Curiel Orosco, Gerardo José. **“Arquitectura para la Edición de Mensajes de Texto en Dispositivos Móviles Utilizando Reconocimiento Automático de Voz ”**. Trabajo Especial de Grado. La Universidad del Zulia. Facultad Experimental de Ciencias. División de Programas Especiales. Maracaibo, Venezuela. 2012. 18p.

RESUMEN

El objetivo de esta investigación consistió en desarrollar una arquitectura para la edición de mensajes de texto utilizando reconocimiento automático de voz, que permita redactar correos y mensajes de texto de una manera mucho más rápida y eficiente, además de permitir una interacción intuitiva y accesible para invidentes y personas que no han aprendido a usar esta tecnología para comunicarse. La metodología para la elaboración del sistema es un modelo de desarrollo por fases siguiendo los lineamientos de los modelos de procesos incrementales. Principalmente, se revisó la literatura sobre motores de reconocimiento automático de voz determinando que Pocketsphinx, herramienta para reconocimiento continuo de voz en tiempo real, resultó la mas adecuada a los requerimientos de la investigación. Se diseñó y construyó la arquitectura del lado del servidor que procesaría las peticiones de reconocimiento de voz. Seguidamente, se desarrolló una interfaz móvil que se comunica con el componente servidor para enviar peticiones de reconocimiento de voz, caracterizada por ser intuitiva y extensible, capaz de asistir en el proceso de redacción de mensajes de texto usando reconocimiento de voz.

Palabras claves: Tecnología Inalámbrica, Móvil, Reconocimiento de Voz, Interfaces, Arquitectura

Correo electrónico: gcuriel@gmail.com

Curiel Orosco, Gerardo José. **“Arquitectura para la Edición de Mensajes de Texto en Dispositivos Móviles Utilizando Reconocimiento Automático de Voz”**. Trabajo Especial de Grado. La Universidad del Zulia. Facultad Experimental de Ciencias. División de Programas Especiales. Maracaibo, Venezuela. 2012. 18p.

ABSTRACT

The objective of this research is to develop a prototype interface for SMS editing using automatic speech recognition that allows writing emails and text messages quickly and efficiently, and allows a much more intuitive and accessible interaction for blind people and people who haven’t learned how to use this technology to communicate. The system development methodology is a phased development model following the guidelines of incremental process models. An initial documental research was carried out to select the automatic speech recognition engine, which pointed to Pocketsphinx, a real-time continuous speech recognition system, as the most suited according to this investigation’s requirements. A server-side architecture for speech recognition was designed and developed. Next, an intuitive, extensible mobile interface for sending speech recognition request to the server component was created, capable to assist in the SMS editing process using speech recognition.

Key Words: Wireless Technology, Mobile, Speech Recognition, Interfaces, Architecture

Email: gcuriel@gmail.com

ÍNDICE DE CONTENIDO

Pág.

ÍNDICE DE GRÁFICOS

Pág.

INTRODUCCIÓN

El mundo de las telecomunicaciones ha ofrecido desde sus comienzos una serie de beneficios tecnológicos que se han aprovechado correctamente; sin embargo, con el transcurso del tiempo han surgido nuevas necesidades y nuevos requerimientos, entre ellos la movilidad y flexibilidad en las comunicaciones.

Actualmente los mercados y las tecnologías se encuentran en un cambio constante, las comunicaciones móviles han sido clave en esta revolución comercial y estratégica, trayendo movilidad a las comunicación en dos vías.

Uno de los métodos de comunicación que presentan los dispositivos móviles, complementaria a la comunicación vía voz, son los mensajes de texto. Sin embargo, la interacción con el teléfono para utilizar este método es ineficiente y lenta. Las empresas fabricantes de teléfonos tampoco han desarrollado mejoras o alternativas de interacción para la composición de mensajes de texto.

Afortunadamente, la lingüística computacional ha generado desarrollo científico teórico y aplicado que pretende incorporar habilidades lingüísticas a las computadoras, permitiendo la comunicación entre maquinas y personas a través de medios mas naturales, por ejemplo mediante una comunicación oral. Particularmente, los reconocedores de voz permiten a un computador transformar automáticamente una entrada de voz en su correspondiente forma textual, efectivamente posibilitando el desarrollo de un método alternativo de entrada de información a un dispositivo móvil, mas natural y lo mas parecida posible a como lo haría un ser humano.

Esta investigación se encuentra estructurada en siete capítulos, conformados cada uno de ellos de la siguiente manera:

El Capítulo I identifica el problema objeto de estudio mediante su debido planteamiento y señala los objetivos, justificación y alcance establecidos en esta investigación.

El Capítulo II está conformado por los antecedentes y los fundamentos teóricos sobre los cuales se asentó la investigación, ofreciendo de esta manera una mejor comprensión del tema abordado.

El Capítulo III describe la metodología empleada para el desarrollo de la investigación, así como los pasos para dar respuesta a los objetivos planteados.

En el Capítulo IV se definen las herramientas empleadas para la ejecución del sistema objeto de la investigación, así como los criterios de selección respectivos y los métodos empleados más sobresalientes.

El Capítulo V muestra la arquitectura del lado del servidor, los detalles de implementación y la integración con el motor de reconocimiento de voz. El Capítulo VI muestra la interfaz de usuario del sistema, gráfica y funcionalmente según su interacción con el usuario.

El Capítulo VII muestra el análisis de los resultados de la investigación mediante las pruebas efectuadas a las herramientas y a la arquitectura en general.

Finalmente, se plantean las conclusiones y recomendaciones obtenidas al culminar esta investigación, brindando la información y aportes necesarios a considerarse en investigaciones similares efectuadas a futuro.

CAPÍTULO 1
El Problema

1. Planteamiento del Problema

El uso de dispositivos móviles revolucionó la manera en que la humanidad se comunica, haciendo posible el realizar y recibir llamadas en cualquier lado, conectando y acercando a las personas cada vez más. Para muchos, en un principio, las llamadas por dispositivos móviles estaban fuera de su alcance, por el alto costo de esta nueva tecnología, en comparación con las tarifas que llevaban las compañías de telefonía fija. Es por esto que con la invención del envío de mensajes de texto, un formato reducido de comunicación, una creciente población adquirió teléfonos móviles para tener todos los beneficios de la movilidad de los celulares pero con bajo costo.

Sin embargo, la interacción necesaria para el envío de un mensaje de texto es ineficiente. La redacción en los teclados pequeños de los teléfonos actuales es inherentemente lenta e inconveniente, además, es una tarea que requiere de toda la atención del usuario; por ejemplo, crear un mensaje mientras se conduce un automóvil se convierte en una tarea extremadamente peligrosa, ya que el conductor podría no prestar toda la atención a la vía mientras redacta el mismo (1).

Las empresas productoras de teléfonos móviles y software, aún cuando han desarrollado innovaciones en el área de interfaz de usuario de estos dispositivos como el marcado rápido y discado de voz, se han quedado atrás en el tan popular envío de mensajes de texto, haciendo que la manera de relacionar al usuario con el dispositivo no haya mejorado desde que el mismo salió el mercado.

Con la realización de este trabajo se buscó el desarrollo de una interfaz y arquitectura alternativa de redacción de mensajes de texto, utilizando reconocimiento de voz, que reduzca al mínimo o incluso elimine la interacción necesaria del usuario para la redacción y envío de mensajes de texto utilizando dispositivos móviles.

Se requirió que la arquitectura fuese robusta y el servicio expuesto fuese fácil de consumir por dispositivos móviles, los cuales actualmente no tienen la capacidad de procesamiento necesaria para llevar a cabo el proceso de reconocimiento de voz por si solos.

2. Alcance del Problema

En investigación estuvo delimitada a la creación de una arquitectura para la redacción de mensajes de texto usando reconocimiento automático de voz y un componente móvil, el cual consume el servicio de reconocimiento de voz ofrecido por esta arquitectura.

3. Objetivos

3.1. Objetivo General

Construir una arquitectura para la edición de mensajes de texto en dispositivos Móviles utilizando reconocimiento automático de voz.

3.2. Objetivos Específicos

- Revisar literatura acerca del tema y aplicaciones existentes para el reconocimiento de voz e interacción con tales sistemas.
- Realizar un análisis comparativo de los motores de reconocimiento automático de voz aplicables.
- Seleccionar la herramienta de reconocimiento automático de voz.
- Diseñar la arquitectura del sistema.
- Diseñar la interfaz gráfica del sistema.
- Implementar el protocolo de comunicación entre la interfaz gráfica y el reconocedor automático de voz.
- Probar la funcionalidad de la aplicación.

4. Justificación de la Investigación

El interés original que motivó esta investigación es el de darle capacidades de reconocimiento de voz a sistemas de lingüística computacional desarrollados en los centros de investigación de La Universidad del Zulia. Estos esfuerzos de investigación requieren de un sistema de procesamiento de emisiones de voz con la capacidad de integración con múltiples tipos de interfaces de usuario, como por ejemplo un dispositivo móvil e incluso otros sistemas lingüísticos.

La capacidad de respuesta textual basada en una entrada de habla proporcionaría otra forma de comunicación entre el dispositivo de cómputo y el hombre que podría ser útil incluso para personas con necesidades especiales (invidentes) y para facilitar en general el uso de tecnología, sustituyendo las interfaces gráficas basadas en dispositivos físicos de entrada por interfaces naturales basadas en voz.

Entre los beneficios que ofrecen sistemas de este tipo están la creación de sistemas de redacción de texto, sistemas del tipo “asistente personal”, sistemas tipo kioscos de información interactivos, interfaces de usuario para personas invidentes, entre otros.

La integración de reconocimiento de voz en sistemas de lingüística computacional permite aumentar la interactividad con el usuario y mejorar la accesibilidad, consiguiendo que el contenido e incluso las órdenes sean reconocidas, lo que posibilita la construcción de sistemas con los que los usuarios pueden interactuar de manera natural.

5. Delimitación de la investigación.

La investigación se llevó a cabo en la Unidad de Lenguajes y Modelos Computacionales de la Facultad Experimental de Ciencias de la Universidad del Zulia y su duración fue de 16 meses. Se inició en el mes de noviembre de 2011 y culminó en marzo de 2012. La línea de investigación a la cual estuvo anexa fue: Lingüística Computacional.

CAPÍTULO 2
Marco Teórico

1. Antecedentes de la Investigación

En la actualidad el campo de investigación en torno al tema de los sistemas de reconocimiento automático de voz a nivel internacional tiene más de dos década de historia. Pero la aplicación de estas tecnologías en dispositivos móviles es de reciente data. Entre estas, se pueden mencionar los siguientes:

- Google Voice Search

El producto Search by voice (2) (en español: “búsqueda por voz”) es la tecnología desarrollada por Google para realizar búsquedas web en PC y teléfonos móviles. La premisa de Voice Search es habilitar al dispositivo usado para buscar datos basados en una entrada de búsqueda en forma de una emisión de voz, que describe el parámetro de búsqueda.

El servicio está disponible en dispositivos móviles a través de una aplicación nativa, disponible para el sistema operativo iOS y para Android. En el caso de las PC, está disponible a través de su navegador web Google Chrome, como un elemento html de entrada con la característica habilitada, usando la clase x-webkit-speech. Actualmente, no tiene soporte para español.

El esquema de trabajo de este producto es basado en un servicio web, expuesto siguiendo la metodología REST para exposición de recursos en internet, que recibe la voz proveniente de los dispositivos capaces de enviar audio y retorna en texto la hipótesis de reconocimiento. El corpus de optimizado para interpretar frases o términos usualmente usados en el contexto de búsquedas de internet.

Dado la naturaleza privativa de la solución, no se dispone de mayor información sobre la arquitectura de reconocimiento que usa Google Voice Search.

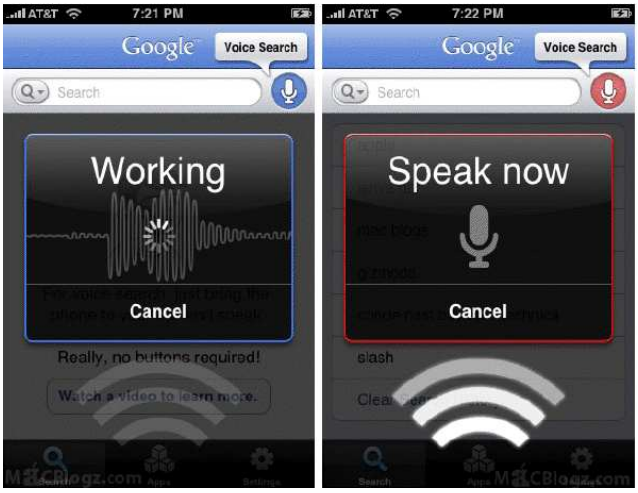

Figura 1. Google Voice. Fuente: Google Inc. (2012)

- Siri

Siri es una aplicación asistente personal para el sistema operativo móvil de Apple, iOS. La aplicación utiliza reconocimiento de voz y posterior procesamiento de lenguaje natural para responder preguntas, hacer recomendaciones y realizar acciones mediante la delegación de las solicitudes a un conjunto de servicios web que va en aumento. El esquema de trabajo de Siri está basado en un servicio web que, de igual manera que Google Voice Search, envía la emisión de voz a un servidor para su reconocimiento.

Siri utiliza conocimiento resultado de más de 40 años de investigación fundada por DARPA a través del Centro de Inteligencia Artificial de SRI International, una organización originada dentro de la Universidad de Stanford.

Actualmente no tiene soporte para español; solo dispone de inglés, alemán, francés y japonés como idiomas de reconocimiento (3).

Figura 2. Siri. Fuente: Apple Inc. (2012)

2. Bases Teóricas

2.1. Reconocedor de Voz

Son sistemas con tecnología que puede traducir palabras habladas en texto. Algunos sistemas de reconocimiento de voz usan ”entrenamiento“ donde una persona lee secciones de texto dentro del sistema. Estos sistemas analizan las características específicas de la voz de la persona y la utilizan para afinar el reconocimiento, resultando en transcripción más exacta. Existen dos tipos de reconocedores de voz: Son llamados ”independientes del hablante“ los sistemas que pueden reconocer una variedad de hablantes, sin ningún entrenamiento. Este tipo de software generalmente limita el número de palabras en un vocabulario, pero es la única opción realística para ciertas aplicaciones que deban aceptar entrada de voz proveniente de un número grande de usuarios. Por otro lado los sistemas que usan entrenamiento son llamados ”dependientes del hablante“, los cuales solo son capaces de reconocer la voz de los usuarios con los cuales fue entrenado (4).

2.2. Modelo Acústico

Un modelo acústico es un archivo que contiene representaciones estadísticas de cada sonido distinto que compone una palabra; cada una de estas representaciones se le asigna una etiqueta llamada fonema. Este modelo es creado tomando grabaciones de audio del habla y transcripciones de texto asociados, y luego haciendo uso de software para generar las representaciones estadísticas de los sonidos que crea cada palabra (5).

2.3. Diccionario Fonético

Un diccionario fonético es un vocabulario que permite localizar palabras por la forma en la que suenan. Estos diccionarios son útiles cuando la ortografía de una palabra es desconocida. Este pareo no es siempre muy efectivo, pero es práctico la mayoría de las veces (5).

2.4. Modelo del Lenguaje

Un modelo del lenguaje es un archivo que contiene una lista larga de palabras y su probabilidad de ocurrencia. Es usado en aplicaciones de dictado para restringir la búsqueda de palabras. Este define cuál palabra puede seguir qué palabras previamente reconocidas y ayuda a restringir significativamente el proceso de pareo mediante la eliminación de palabras que no son probables. Los modelos de lenguaje más comúnmente usados son modelos de n-grama, los cuales contienen estadísticas de secuencias de palabras (5).

2.5. Corpus de Voz

La idea de un corpus de voz nace del origen mismo de la comunicación en los seres humanos. La creación de registros históricos del habla se remonta a los primeros experimentos de Tomas Edison, que fue quien abrió a la humanidad las puertas de este complejo universo. Los avances tecnológicos del siglo XX progresivamente posibilitaron hacer grabaciones de mayor calidad; inicialmente los elementos magnetofónicos protagonizaron la industria de los registros sonoros, sin embargo durante las últimas décadas se han desarrollado técnicas de digitalización que han incrementado notablemente la calidad de las grabaciones, y han posibilitado el uso del procesamiento digital de señales; estas técnicas derivaron (de entre muchos otros experimentos) en el desarrollo de tecnologías del habla, como por ejemplo reconocimiento de la voz (5).

2.6. HTTP

Hypertext Transfer Protocol o HTTP (en español: protocolo de transferencia de hipertexto) es el protocolo usado en cada transacción sobre la World Wide Web. Es un protocolo orientado a transacciones y sigue el esquema petición-respuesta entre un cliente y un servidor. Al cliente

que efectúa la petición (un navegador web o un spider) se lo conoce como ”user agent“ (agente del usuario). A la información transmitida se la llama ”recurso“ y se la identifica mediante un localizador uniforme de recursos (por sus siglas en ingles: URL). Los recursos pueden ser archivos, el resultado de la ejecución de un programa, una consulta a una base de datos, la traducción automática de un documento, etc (6).

2.7. Códec

Códec es la abreviatura de codificador-decodificador. Describe una especificación desarrollada en software, hardware o una combinación de ambos, capaz de transformar un archivo con un flujo de datos (en ingles: stream) o una señal. Los códecs pueden codificar el flujo o la señal (a menudo para la transmisión, el almacenaje o el cifrado) y recuperarlo o descifrarlo del mismo modo para la reproducción o la manipulación en un formato más apropiado para estas operaciones. Los códecs son usados a menudo en videoconferencias y emisiones de medios de comunicación (7).

2.8. JSON

JSON, acrónimo de JavaScript Object Notation, es un formato ligero para el intercambio de datos, basado en texto y diseñado como contenedor para intercambio de datos. Es derivado del lenguaje Javascript, donde es usado para representar estructuras de datos y arreglos asociativos. La simplicidad de JSON ha dado lugar a la generalización de su uso. El formato JSON se emplea habitualmente en entornos donde el tamaño del flujo de datos entre cliente y servidor es de vital importancia (8).

2.9. JavaScript

JavaScript es un lenguaje de programación interpretado, dialecto del estándar ECMAScript. Se utiliza principalmente en su forma del lado del cliente, implementado como parte de un navegador web permitiendo mejoras en la interfaz de usuario y páginas web dinámicas, en bases de datos locales al navegador (9).

2.10. SMS

El servicio de mensajes cortos o SMS (del inglés: Short Message Service) es un servicio disponible en los teléfonos móviles que permite el envío de mensajes cortos (también conocidos como mensajes de texto, o más coloquialmente, textos) entre teléfonos móviles, teléfonos fijos y otros dispositivos de mano (10).

2.11. Smartphone

Un teléfono inteligente (del ingles: smartphone) es un teléfono móvil construido sobre una plataforma de informática móvil, más la capacidad de computación avanzada y conectividad de un teléfono móvil, con la posibilidad de instalar aplicaciones para cualquier uso. El término «inteligente» hace referencia a la capacidad de usarse como un computador de bolsillo, llegando incluso a remplazar a un computador personal en algunos casos (11).