Estudio sobre la navegación en la Web usando voz

Areli Torres González [1], Luis Villaseñor Pineda [2]

[1] Facultad de Ciencias Computacionales, BUAP
Av. San Claudio y 14 Sur
Puebla Pue.
are67@hotmail.com.mx

[2] Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Tonantzintla, Puebla.
villasen@inaoep.mx

Resumen. El presente trabajo muestra un estudio empírico de la interacción hombre-máquina al usar la voz para la navegación en la Web. Para ello se grabaron una serie de experimentos en condiciones controladas, y se realizó un análisis de las transcripciones correspondientes. El análisis realizado hasta ahora se ha enfocado principalmente al rendimiento del reconocedor de voz y posteriormente se extenderá a las implicaciones del uso de la voz en la interacción hombre-máquina. Entre los primeros resultados obtenidos tenemos una tasa de reconocimiento del 76.82% y una tasa de ejecución del 72.74%. También se han identificado un primer grupo de puntos críticos en la interacción vocal, tanto a nivel del proceso de reconocimiento de voz como de la utilidad de la voz al navegar en la Web.

Introducción

Actualmente Internet es una de las herramientas más importantes con las que cuenta el hombre moderno. Proporcionando múltiples productos y servicios de información, tales como el Web (Word Wide Web), en donde es posible la navegación hipertextual y permite el acceso a información en forma de texto, sonido, e imágenes animadas [1]. Gracias a los recientes avances en el reconocimiento automático de voz se ha iniciado el uso comercial de esta tecnología en la navegación en la Web [2]. El presente trabajo muestra un estudio empírico de interacción hombre-máquina al usar la voz para la navegación en Internet. El objetivo general de este estudio fue la evaluación de la aplicación de esta técnica de interacción. Para ello se grabaron una serie de experimentos en condiciones controladas, y se realizó un análisis de las transcripciones correspondientes. Las siguientes secciones presentan la plataforma usada para llevar a cabo los experimentos, la descripción de los mismos, y el análisis de sus transcripciones. Por

Gelbukh, M. Hernández Cruz (Eds.) Avances en la ciencia de la computación en México, CORE-2003, pp. 149–158, 2003.

último, se presentan las ventajas y desventajas que podemos concluir de este estudio, tanto a nivel del proceso de reconocimiento de voz como de la utilidad de voz al navegar en la Web.

2 Descripción de la plataforma de navegación

Para llevar a cabo este estudio se preparó una plataforma de navegación. En esta sección se mencionan algunas de sus principales características. Esto con el fin conocer las condiciones en las cuales se realizaron los experimentos.

Tabla 1 . Comandos vocales para la navegación en Web

NAVEGAR EN LA WEB	
Ir atrás	Detener
Ir adelante	Ir a la página de inicio
DESPLAZARSE POR UNA PÁGINA WEB	
Página abajo	Desplazar hacia abajo
Página arriba	Desplazar hacia arriba
Línea abajo	Detener desplazamiento
Ir al comienzo	Más rápido
Ir al final	
MANEJO DE IMÁGENES (HIPERVINCULO)	
Hacer clic en Imagen	Elegir n (n = 1...20)
Imagen	Ir n abajo (n = 1...20)
COMANDOS GLOBALES	
Pulsar Entrar	Marco siguiente
Cambiar a ventana siguiente	Edición
Cambiar a ventana anterior	Seleccionar Todo
A Dormir	Ir abajo
A Trabajar	Ir al inicio
Archivo	Ir arriba
Guardar como	Marco anterior
Cancelar	

2.1 Materiales

Para efectos de este estudio se limitó la navegación a un sólo sitio web. Se buscó sitio con una estructura fija pero cuyo contenido cambiará día con día. En nuestro caso decidimos utilizar el sitio de La Jornada, un periódico de circulación nacional (La Jornada Virtual http://www.jornada.unam.mx/). Este fue elegido ya que presenta una estructura fija pero cuya información cambia diariamente.

También se hizo uso del sistema de reconocimiento de voz *Dragon Naturally Speaking*. Este paquete comercial incluye un conjunto de comandos orales que nos

permiten controlar Internet Explorer y seguir los vínculos con la voz. Además nos ofrece un conjunto de herramientas para la creación de nuevos comandos *orales* [3][4], lo que nos permite crear nuestros propios comandos para así poder automatizar una serie de tareas adicionales sobre el Explorador. En la tabla 1 se enlistan los comandos que un usuario puede usar para navegar en la Web y en la Tabla 2 se muestran los comandos adicionales que se crearon. En particular, se crearon comandos orales para controlar los estados de la ventana del Explorador y poder seleccionar marcos.

La recopilación de los datos para este estudio se hizo usando dos herramientas de apoyo: *Techsmith Camtasia Recorder*, la cual nos permitió capturar el video de las sesiones experimentales; y *Teleport Pro*, una herramienta que nos permitió bajar todo un sitio web y así trabajarlo de manera local, evitando largos tiempos de espera.

Tabla 2. Comandos Adicionales

MANIPULACIÓN DE VENTANAS	
Maximizar	Restaurar
Minimizar	Cerrar
MANEJO DE MARCOS	
Seleccionar Marco	

3 Descripción de los Experimentos

Para lograr el objetivo del estudio de la navegación en la Web mediante la voz se diseñaron una serie de experimentos cuyo propósito principal era obtener un conjunto de resultados que nos permitieran concluir las ventajas y desventajas de este tipo de navegación, así como las características que presenta este tipo de interacción [6].

Los experimentos fueron realizados por 10 personas. A cada una de ellas se le asignó una tarea –la misma tarea para todas– a realizar navegando en el sitio de la Jornada. A continuación se detallan cada uno de los pasos del experimento.

3.1 Etapas de los Experimentos

1. Explicación breve sobre la prueba y sus objetivos. Se explicó brevemente al usuario el objetivo y los alcances de la prueba. Este punto era importante pues se quería dejar claro que él no sería el evaluado sino que con su ayuda se deseaba evaluar al sistema.
2. Entrega del 1er cuestionario. Previamente a la realización de la tarea, a cada usuario le fue entregado un breve cuestionario. Se solicitaron sus datos

personales e información de control de carácter general, para obtener el de cada uno de ellos.

3. Adecuación del sistema de reconocimiento al usuario. En esta etapa, siguieron los pasos para adaptar Dragon a la voz del usuario, esto con el fin obtener la mejor taza de reconocimiento posible. A través del *entrenamiento sistema* se adaptan los modelos acústicos a las características específicas cada hablante; como timbre, tono, volumen y acento. Dicho entrenamier consistió en leer en voz alta y clara parte de un libro o una historia dura aproximadamente cinco minutos.
4. Explicación y uso de comandos vocales por medio de un ejemplo. A usuario se le explicaron cada uno de los comandos que podía utilizar navegar sobre *La Jornada Virtual.* Se les proporcionó una lista con todos comandos y sus especificaciones (ver la Tabla 1 y 2). Después se llevó a un ejemplo sencillo con el fin de mostrarle a cada usuario como usar comandos vocales. El ejemplo consistió en:
   - Buscar en la sección de Espectáculos, dentro del sitio de *La Jorna Virtual,* un reportaje sobre Fernando Sariñana y guardarlo en disco.
   - Ir a *La Jornada de Oriente* y visitar la sección de Puebla, al termi regresar a *La Jornada Virtual.*
5. Elaboración de los ejercicios. Para ello se seleccionó previamente un conju de ejercicios que el usuario realizaría haciendo uso de la mayoría de comandos de navegación. Cada ejercicio incluyó una serie de pregun asociadas, las cuales deberían ser contestadas conforme se llevará a cabo ejercicio. A continuación se enlistan los ejercicios solicitados:
   - Entra a la pagina de La Jornada, elige un encabezado de tu interés y léelo.

     ¿Cómo se titula el encabezado?
     ¿De cuantos párrafos consta dicho encabezado?
   - Busca la sección de cartones y contesta lo siguiente:

     ¿Cuál es el título del cartón que más te agrado?
     ¿Quién es el autor de dicho cartón?
   - Entra a la sección de CineGuía, de la cartelera mostrada:

     ¿Qué película te gustaría ver?
     ¿Qué clasificación es?
     ¿Quiénes son los protagonistas?
   - Elige una sección del periódico al azar. De dicha sección menciona la noticia artículo más impactante y guárdalo en disco.

     Título de la Noticia/Artículo
   - Visita brevemente el sitio de la Jornada de Oriente y regresa a la página Jornada al terminar.

     ¿Qué estados conforman la Jornada de Oriente?

6. Entrega del 2do cuestionario. Al terminar los ejercicios se solicita la opinión general del usuario a cerca de la navegación por medio de voz así como la descripción de los posibles problemas que enfrentó.

Tabla 3. Tiempos de Grabación

USUARIOS	TIEMPO DE GRABACIÓN
usuario 1	16.57 min.
usuario 2	11.10 min.
usuario 3	09.01 min.
usuario 4	04.57 min.
usuario 5	18.48 min.
usuario 6	27.49 min.
usuario 7	17.47 min.
usuario 8	16.59 min.
usuario 9	13.16 min.
usuario 10	12.30 min.
Tiempo promedio de grabación	14.673 min.

4 Transcripción de las grabaciones

El primer paso antes de iniciar propiamente el análisis de los resultados fue la transcripción de las grabaciones. Durante cada experimento se grabó un video para mantener un registro completo de todas los pasos realizados por el usuario para llevar a cabo la tarea asignada. La Tabla 3 muestra los tiempos de grabación por usuario.

LO QUE SE DIJO	LO QUE SE RECONOCIO	LO QUE SE EJECUTO
Ir atrás	Ir atrás	Cargo la página anterior

Fig. 1. Transcripción de un comando

La transcripción consiste en llevar a texto cada una de las intervenciones del usuario junto con los resultados obtenidos. En este caso se recopilaron tres tipos de resultados por cada intervención:

- *Lo que se dijo.* Se transcribió ortográficamente la elocución pronunciada por usuario.
- *Lo que se reconoció.* Se transcribió el texto reportado por Dragon resultado del proceso de reconocimiento de voz. En caso de que el programa no reporta ningún texto reconocido colocamos un guión (-).
- *Lo que se ejecutó.* Se describió la acción que realizó el navegador. En caso de que el programa no llevaba a cabo ninguna acción colocamos un guión (-).

De esta manera, se creó una tabla por cada sesión. Las tablas estaban constituidas por tres columnas representando los tres tipos de resultado para cada intervención.

5 Análisis de las transcripciones

Tomando como referencia las transcripciones de cada usuario se realizó su análisis, mediante el estudio de dos relaciones posibles: *lo que se dijo-lo que se reconoció* *lo que se dijo – lo que se ejecutó.*

Es importante recordar que este estudio está en proceso y en este trabajo presentan los resultados iniciales que se enfocan principalmente en comportamiento del reconocedor de voz. Las dos relaciones anteriores tratan capturar este comportamiento.

5.1 Lo que se dijo-lo que se reconoció

Para esta relación se trabajó con las columnas uno y dos de las transcripciones. Las frases dichas por el usuario se clasificaron en tres tipos:

- *Comandos.* Las elocuciones que ordenaban a Dragon la ejecución de una acción.
- *Vínculos.* Las frases dichas por el usuario que correspondían a vínculos entre páginas.
- *Otros.* Son las frases que no corresponden a ninguno de los casos anteriores. trata de frases que el usuario intuitivamente creó para designar una acción; aquellos vínculos que aparecían sobre la página web como si se tratará de textos pero que en realidad eran imágenes.

Cada una de las intervenciones del usuario fue clasificada usando estos criterios. Posteriormente se obtuvieron el número de intervenciones totales del usuario, como el número de intervenciones reconocidas y las intervenciones reconocidas errores. Un error es cuando el proceso de reconocimiento de voz substituía la orden

pronunciada por el usuario con alguna otra orden válida. Es decir, Dragon reportó algo parecido a lo dicho por el usuario pero que no es lo inicialmente solicitado.

La tabla 4 muestra los resultados de la relación 1. En ella podemos observar que de las 686 frases dichas entre los diez usuarios Dragon reconoció 556 (81.05%) de las cuales sólo 527 (76.82%) fueron órdenes solicitadas expresamente por el usuario. Es decir, 29 frases fueron mal interpretadas.

Tabla 4. Relación 1: Lo que se dijo – lo que se reconoció

RELACIÓN 1:	LO QUE SE DIJO	LO QUE SE RECONOCIO	
USUARIOS	DIJO	SE RECONOCIO	SE RECONOCIO SIN ERRORES
usuario 1	104	65	57
usuario 2	58	53	51
usuario 3	45	34	34
usuario 4	33	30	30
usuario 5	82	60	58
usuario 6	94	89	83
usuario 7	62	47	46
usuario 8	84	65	58
usuario 9	60	53	51
usuario 10	64	60	59
Número de Frases totales:	686	556	527

Tabla 5. Relación 2. Lo que se dijo – lo que se ejecutó

RELACION 2:	LO QUE SE DIJO	LO QUE SE EJECUTO	
USUARIOS	DIJO	SE EJECUTO	SE EJECUTO SIN ERRORES
usuario 1	104	57	53
usuario 2	58	54	51
usuario 3	45	31	31
usuario 4	33	30	30
usuario 5	82	53	53
usuario 6	94	77	76
usuario 7	62	42	42
usuario 8	84	55	54
usuario 9	60	53	52
usuario 10	64	57	57
Número de Frases totales:	686	509	499

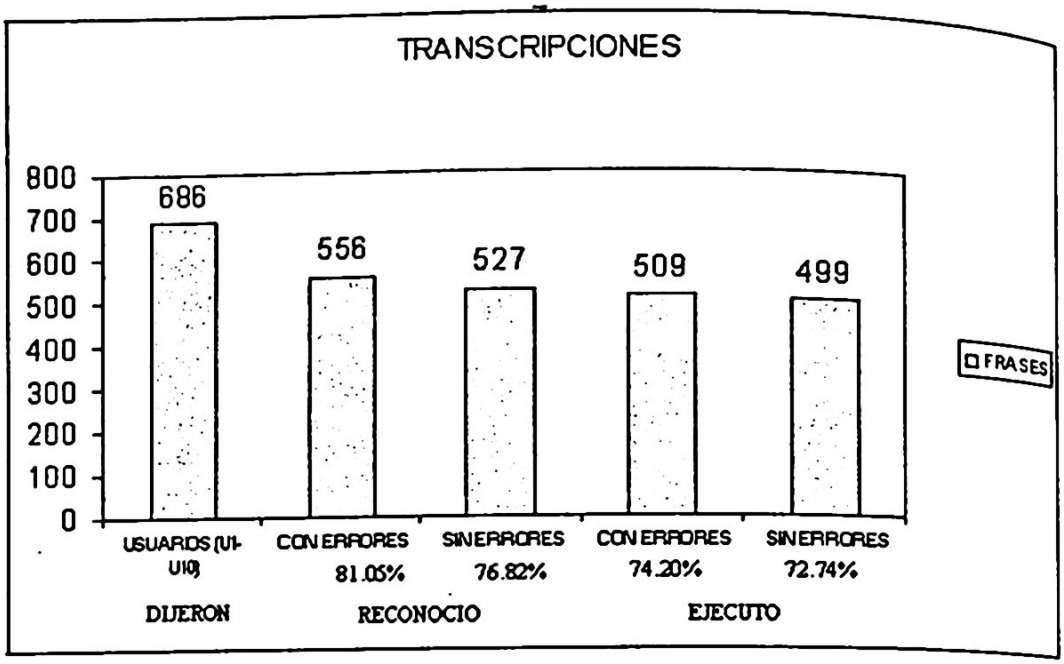
5.2 Lo que se dijo – lo que se ejecutó.

En esta relación se tomaron como referencia la primera y tercera columna de las transcripciones de cada usuario. De igual forma como en la primera relación, las dos columnas se clasificaron en tres tipos: comandos, vínculos y otros.

De igual manera, se calcularon para cada usuario el número de veces que lo ejecutó, no lo ejecutó o bien se ejecutó otra cosa. Los resultados obtenidos se muestran en la tabla 5. De esta tabla notamos que de las frases dichas por los usuarios se ejecutaron en total 74.20% y un 72.74% sin errores.

Así podemos concluir que la tasa de ejecución durante estos experimentos fue un 72.74%. Es interesante notar, que tanto en el proceso de reconocimiento como ejecución existen sustituciones realizadas por el sistema. En resumen los resultados totales se presentan en la tabla 6.

Tabla 6. Resultado del análisis de las transcripciones



5.3 Perfil de los usuarios

De los datos recabados a través de los cuestionarios es importante recalcar siguientes:

- de la muestra de diez usuarios, ocho eran hombres y dos mujeres.
- siete de los usuarios tenían estudios en el área de Ciencias Computacionales, uno en Lingüística, uno en Matemáticas y uno en Diseño Gráfico
- el rango de edad fue de 21 años a 27 años en general, excepto un usuario tenía 54 años.
- el 40% de los usuarios había tenido contacto previo con sistemas reconocimiento automático de voz (p.e. Sistemas de dictado)
- el 70% opinó, después de realizada la prueba, que el uso de la voz facilita navegación.

6 Resultados preliminares

Dos tipos de conclusiones se desprenden de este estudio hasta este momento. primeras son conclusiones con respecto al sistema de reconocimiento de voz (en caso en particular, Dragon Naturally Speaking), pero que pueden aplicarse cualquier sistema de reconocimiento automático de voz. Un segundo grupo

conclusiones presentan observaciones sobre el uso de la voz en este tipo de situación.

Sobre el desempeño del sistema de reconocimiento de voz tenemos los siguientes comentarios:

El medio ambiente de trabajo: El ruido ambiental es un serio problema, se debe buscar disminuirlo y si es que lo hubiera, que éste sea constante.

Los locutores: El uso de la voz, independientemente de la aplicación, depende del tiempo dedicado en el entrenamiento del sistema de reconocimiento. Mientras más se adapte el sistema a nuestra voz la tasa de reconocimiento será mejor. Esto implica cuidados e interés por parte del usuario que en ocasiones serán un obstáculo.

El sistema de cómputo utilizado: Es importante contar con un equipo potente, de acuerdo a los requerimientos de Dragon Naturally Speaking, para contar con un tiempo de respuesta adecuado.

El modelo de lenguaje: El sistema de reconocimiento inclina su reconocimiento por un modelo de lenguaje predeterminado. Esa es la principal explicación de los errores introducidos por el reconocedor al sustituir órdenes válidas correctamente pronunciadas por el usuario. El uso de frases cortas o la presencia de un modelo de lenguaje que prefiere cierto tipo de construcción gramatical sobre otro son las principales razones.

A pesar de no haber terminado el estudio respecto al uso de la voz en la navegación podemos presentar la siguiente lista de comentarios preliminares, contrastándolos en ventajas y desventajas.

Entre las ventajas de este modo de interacción podemos reportar las siguientes, tanto por ser observadas directamente, como remarcadas por los usuarios en los cuestionarios de control.

- *Realizar diferentes actividades*: El usuario tendrá la libertad de usar sus manos para efectuar otra tarea mientras navega, siempre y cuando tenga un conocimiento completo de la estructura de la página o sitio web a consultar.
- *Cómoda y de fácil manejo.*
- *Útil para personas con capacidades diferenciadas.* (p.e. con problemas motores en sus extremidades superiores)

Entre las desventajas deseamos señalar las siguientes:

- *Uso de un sub-lenguaje:* El usuario debe memorizar comandos específicos lo que restringe fuertemente el uso del lenguaje oral, e impone restricciones a la "naturalidad" esperada de este tipo de comunicación.
- *División entre la tarea de navegación y el contenido*: El esquema de navegación estudiado pretende imponer una clara división entre la tarea de navegación y el contenido de las páginas. Al navegar usando la voz esta división desaparece. Varios usuarios presentaron en más de una ocasión, y a

pesar de haber recibido una demostración de la navegación, preferencia por uso de comandos que se relacionaban directamente con el contenido.

7 Conclusiones

El presente estudio es parte de una serie de esfuerzos para poder delimitar claramente cuales son los fenómenos propios de la comunicación oral humana que deben presentarse en la comunicación oral hombre-máquina para que ésta sea realmente útil [5] [7].

Este trabajo es de particular importancia pues estudia la integración de la voz un ambiente diseñado y desarrollado para la presentación gráfica de información. Una situación donde la intervención humana podría hacerse de manera verbal pero la presentación de la respuesta sería fundamentalmente gráfica. Bajo estas condiciones la comunicación oral hombre-máquina no presenta ningún paralelismo con la comunicación oral humana.

Agradecimientos

Los autores agradecen al CONACYT por el apoyo financiero para la realización este estudio (No. Ref 31128A). En particular, la beca de tesis de licenciatura recibida por el primer autor para el desarrollo de este trabajo. Los autores también agradecen el apoyo del Laboratorio de Tecnologías del Lenguaje del INAOE.

Referencias

1. Hahn, Harley. *Internet, Manual de Referencia.* - McGraw-Hill, 1994.
2. Bernal Bermúdez Jesús, Bobadilla Sancho Jesús, Gómez Vilda Pedro: *Reconocimiento Voz y Fonética Acústica.* –Alfaomega Grupo Editor, 2000.
3. Dragon Naturally Speaking: www.dragonsys.com.
4. Dragon Naturally Speaking: *Creating Voice Commands.* –Learnout & Hauspie, September 2000.
5. David N. Chin & Martha E. Crosby. Introduction to the Special Issue on Empirical Evaluation of User Models and User Modeling Systems, *User Modeling and User-Adapted Interaction*, Vol. 12, No. 2-3, Kluwer, 2002
6. Diane J. Litman & Shimei Pan. Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction*, Vol. 12, No. 2-3, Kluwer, 2002.
7. Lynnette Hirshman & Henry S. Thompson. Overview of Evaluation in Speech Natural Language Processing. *Survey of the State of the Art in Human Language Technology.* Ronald Cole (ed.) National Science Foundation, 1995.