ISSN: 1692-7257 - Volumen 1 – Número 45 - 2025

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i45.2963

Acceso a Internet por Comandos de Voz: Aplicación de Navegación para Facebook, Gmail y Chrome

Internet Access by Voice Commands: Navigation Application for Facebook, Gmail and Chrome

MSc. José Hernando Mosquera de la Cruz[1], PhD. Humberto Loaiza Correa[1]
PhD. Sandra Esperanza Nope Rodríguez[1]

[1] ***Universidad del Valle,*** *Facultad de Ingeniería, Escuela de Ingeniería Eléctrica y Electrónica, Cali, Colombia.*

Correspondencia: *sandra.nope@correounivalle.edu.co*

Recibido: *14 junio 2024.* ***Aceptado:*** *20 diciembre 2024.* ***Publicado:*** *01 enero 2025.*

Cómo citar: *J. H. Mosquera Cruz, H. Loaiza Correa, y S. E. Nope Rodríguez, «Acceso a Internet por Comandos de Voz: Aplicación de Navegación para Facebook, Gmail y Chrome», RCTA, vol. 1, n.º 45, pp. 183–194, ene. 2025. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/2963*

Resumen: Se presenta un sistema para navegar por Internet mediante comandos de voz. La herramienta implementada permitió el control verbal de las aplicaciones Google Chrome, Gmail y Facebook. Las pruebas se realizaron con un grupo de 33 personas con diferentes experiencias de navegación por Internet compuesto por adultos jóvenes, adultos mayores y personas con discapacidad motriz. Cada una de las aplicaciones se probó por separado mediante diálogos guiados con comandos de voz y dictados. En las pruebas del sistema de reconocimiento de voz se utilizaron 2871 comandos de voz y 594 dictados, observándose un mejor resultado de los comandos de voz en la aplicación Facebook y de los dictados en la aplicación Google Chrome. Se obtuvo una media general del 84,69% con una desviación estándar del 6,45% para el reconocimiento de comandos de voz, y del 74,63% con una desviación estándar del 2,75% para el reconocimiento de dictados.

Palabras clave: Interacción Humano-Computador, Tecnologías de la Información y la Comunicación, Navegación en Internet, Reconocimiento del Habla.

Abstract: A system to navigate the internet using voice commands is presented. The implemented tool allowed verbal control of Google Chrome, Gmail and Facebook applications. The tests were conducted on a group of 33 people with different experiences browsing the Internet composed of young adults, older adults and people with motor disabilities. Each of the applications was tested separately using guided dialogues with voice commands and dictations. In the speech recognition system's tests, 2871 voice commands and 594 dictations were used, observing a better result for voice commands in the Facebook application and dictation in the Google Chrome application. A general average of 84.69% with a standard deviation of 6.45% was obtained for the recognition of voice commands, and 74.63% with a standard deviation of 2.75% for the recognition of dictations.

Keywords: Human-Computer Interaction, Information and Communications Technology, Internet Navigation, Speech Recognition.

1. INTRODUCCIÓN

Desde la creación de los computadores, los métodos de interactuar con las aplicaciones (teclado y mouse) no fueron diseñados considerando la manera natural en la que interactúan los humanos, por lo que, sin intención, se excluyen personas con limitaciones motrices e incluso de analfabetas tecnológicos. En contraste, las interfaces humano-máquina (HMI, Human-Machine Interface) buscan una interacción más natural [1]. Dado que el habla es la forma más rápida y natural de comunicación entre humanos [2], se constituye en una opción de interacción humano-máquina, asociando secuencias de palabras con comandos de máquina bajo condiciones de calidad del audio desafiantes, como el ruido ambiental, la calidad de la dicción, el acento, la entonación, el timbre, el volumen, y otros factores [3][4]; sin embargo, permitiría sobrepasar las barreras que imponen las interfaces tradicionales [5].

La presencia de computadores e internet en ambientes académicos, laborales y personales ha modificado las formas de comunicación entre humanos [6]. El uso masivo de aplicaciones como Google Chrome, Gmail, Facebook, entre otras, han permeado en mayor o menor medida dichos ámbitos; en particular, en condiciones de confinamiento social por condiciones de bioseguridad.

Los sistemas de reconocimiento y síntesis de voz ya han sido integrados comercialmente en dispositivos móviles y asistentes personales virtuales. Entre los primeros se destacan SIRI de Apple [7], CORTANA de Microsoft [8], GOOGLE ASSISTANT [9], ALEXA [10] y BIXBY de Samsung [11], de Apple, Microsoft, Google y Samsung, respectivamente. Otros desarrollos diseñados para ayudar a los usuarios en sus tareas diaria y proveer un acceso fácil a datos estructurados, servicios web y aplicaciones personales, se encuentran CERENCE DRIVE [12] para asistencia de automóviles, o ASTRO [13]. Estos asistentes requieren la combinación de tecnologías de reconocimiento de voz, procesamiento natural del lenguaje, gestión de diálogo, generación de lenguaje y síntesis de texto a voz [14][15][16][17].

Se distinguen tres tendencias de asistentes personales virtuales comandados por voz. La primera orientada a la domótica [18][19][20][21][22][23]. La segunda orientada a la navegación de vehículos [24][25][26][27][28][29][30], que incluye, además de automóviles, sillas de ruedas y aviones. La tercera, en la que se enmarca el presente trabajo, está orientada al acceso de las tecnologías de la información y la comunicación (TIC), para acceder a información de interés diario tales como distados y corrección de texto [31], clima y noticias [32], y manejo del calendario [33].

En este trabajo se presenta un sistema para la navegación en internet mediante comandos de voz en Facebook, Gmail y Chrome, permitiendo la realización de dictados y la síntesis de señales de voz para retroalimentar al usuario. A continuación, se presentan la metodología, la descripción del sistema, pruebas y conclusiones.

2. METODOLOGÍA

Se seleccionaron las aplicaciones de navegación más reconocidas socialmente. Luego, se identificaron las opciones más usadas dentro de ellas para asociarlas con comandos de voz. En la identificación de comandos de voz, se integró un subsistema de reconocimiento, y, para que el usuario conozca el estado del sistema, se introdujo un subsistema de síntesis de voz. Finalmente, se implementaron interrupciones vía sistema operativo, para que se ejecuten las rutinas asociadas a los comandos de voz.

La evaluación de la interface se llevó a cabo basándose en el análisis de dos fuentes de información: encuestas para medir el nivel de satisfacción de los usuarios con los comandos de voz y dictados, y métricas de desempeño en el reconocimiento del habla.

3. DESCRIPCIÓN DEL SISTEMA

El sistema está constituido por seis bloques, como se ilustra en la (Fig. 1). El proceso inicia con la adquisición de la señal de voz del usuario mediante un micrófono conectado el sistema de cómputo. Inmediatamente después, el bloque de reconocimiento del habla procesa el audio capturado y lo convierte en cadenas de texto. El bloque de validación de comandos identifica si el contenido de la cadena de texto corresponde o no a una de las ordenes establecidas en el diccionario de comandos, y envía esta información al bloque unidad de control; finalmente, el sistema realimenta auditivamente al usuario usando el bloque de síntesis de voz. El código fuente del sistema desarrollado se puede consultar en [34].

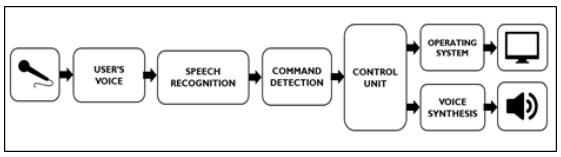

Fig. 1. *Diagrama de bloques del sistema.*
Fuente: *elaboración propia.*

3.1. Adquisición de la Voz del Usuario

La adquisición de la voz del usuario se realizó con el micrófono incorporado en la diadema Logitech G430 [35], en un ambiente semicontrolado con condiciones de ruido de tipo oficina.

3.2. Reconocimiento del Habla

La interfaz propuesta se desarrolló en lenguaje de programación C# bajo el Sistema Operativo Windows 8.1. El sistema de reconocimiento del habla es independiente del locutor, y utilizó las librerías de Bing Speech API [36] que pertenecen a Microsoft Cognitive Services, garantizando la compatibilidad e integración con el resto de algoritmos. El uso de estas librerías requiere una conexión constante a Internet, ya que el reconocimiento del habla y su transcripción se realizan en línea [37]. La arquitectura interna del bloque de reconocimiento de voz se ilustra en el Fig. 2, y consiste de tres modelos: Modelo Acústico, Modelo de Pronunciación y Modelo del Lenguaje.

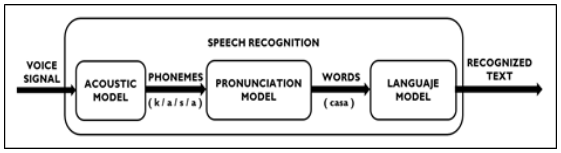

Fig. 2. *Diagrama de bloques del reconocimiento de voz.*
Fuente: *elaboración propia.*

Los modelos que componen el bloque de reconocimiento del habla, se basan en la teoría probabilística [38] para encontrar la secuencia de palabras que más se ajusta a la señal de voz capturada, considerando las distribuciones de probabilidad $P(X|L)$, la pronunciación de los fonemas dentro de un diccionario definido $P(L|W)$, y el lengua la probabilidad de pronunciación de los fonemas dentro de un diccionario definido, y la probabilidad de la organización de las palabras, de modo que sea sintácticamente coherente $P(W)$:

$$argmax\, P(W|X) = argmax\, P(X|L)P(L|W)P(W) \quad (1)$$

3.3. Validación de comandos

Esta parte del sistema, compara el texto reconocido de la voz del usuario, con el diccionario de comandos establecido, que se presenta en la Tabla 1 (segunda columna), y se divide en 3 categorías: comandos genéricos, que son comunes a las tres aplicaciones, y comandos específicos para cada aplicación. Se usó el mismo diccionario que fue propuesto en [39].

Tabla 1: *Comandos de voz, interrupción al sistema operativo y mensajes de síntesis de voz*

	Comando	Teclas interrupción sistema operativo	Mensaje síntesis de voz
Comandos Genéricos	"Abrir Facebook"	Send("{F6}m.facebook.com{ENTER}")	"Abriendo Facebook"
	"Abrir Gmail"	Send("{F6}gmail.com{ENTER}")	"Abriendo Gmail"
	"Abrir Google"	Send("{F6}google.com{ENTER}")	"Abriendo Google, ¿Que deseas buscar?"
	"Salir"	Send("{ALTDOWN}F4{ALTUP}")	"Hasta Pronto"
	"Actualizar"	Send("{F5}")	"Actualizando"
	"Acercar"	Send("{CTRLDOWN}{+}{CTRLUP}")	"Acercando"
	"Alejar"	Send("{CTRLDOWN}{-}{CTRLUP}")	"Alejando"
	"Bajar"	Send("{PGDN}")	"Bajando Página"
	"Subir"	Send("{PGUP}")	"Subiendo Página"
	"Siguiente"	Send("{DOWN}")	"Siguiente Opción"
	"Anterior"	Send("{UP}")	"Anterior opción"
	"Atrás"	Send("{BROWSER_BACK}")	"Volviendo atrás"
	"Insertar Dictado"	Send(*String*)	"Que deseas dictar"
	"Enter" "Entrar" "Aceptar"	Send("{ENTER}")	"Presionando Enter" "Entrando" "Okey"
	"Deshacer"	Send("{CTRLDOWN}z{CTRLUP}")	"Deshaciendo"
	"Escape"	Send("{ESC}")	"Presionando Escape"
Facebook	"Abrir Muro"	Send("{ALTDOWN}1{ALTUP}")	"Abriendo muro"
	"Abrir Notificaciones"	Send("{ALTDOWN}4{ALTUP}")	"Abriendo notificaciones"
	"Abrir Perfil"	Send("{F6}m.facebook.com/me{ENTER}")	"Abriendo perfil"
	"Abrir Mensajes"	Send("{ALTDOWN}3{ALTUP}")	"Abriendo mensajes"
	"Nuevo Estado"	Send("p");	"¿Cuál es tu nuevo estado?"
	"Publicar Estado"	Send("{TAB}"); Send("{ENTER}")	"Publicando Estado"
Chrome	"Nueva búsqueda"	Send("{F6}google.com{ENTER}")	"Que deseas Buscar"
	"Navegar"	Send("{TAB}")	"Navegando Resultados"
	"Descargas"	Send("{CTRLDOWN}j{CTRLUP}")	"Estas son las descargas"
	"Historial"	Send("{CTRLDOWN}h	"Este es el

	Comando	Teclas interrupción sistema operativo	Mensaje síntesis de voz
		{CTRLUP}")	historial"
	"Imprimir"	Send("{CTRLDOWN}p {CTRLUP}")	"Configurando Impresión"
Gmail	"Correos enviados"	Send("g"); Send("t")	"Abriendo mensajes enviados"
	"Abrir Bandeja"	Send("g"); Send("i")	"Abriendo bandeja de entrada"
	"Nuevo correo"	Send("c")	"Creando un nuevo correo, ¿a quién lo vas a enviar?"
	"Agregar Asunto"	Send("{TAB}")	"¿Cuál es el asunto del correo?"
	"Agregar Mensaje"	Send("{TAB}")	"Cuál es el mensaje de correo"
	"Enviar Correo"	Send("{TAB}"); Send("{ENTER}")	"Enviando Correo"

Fuente: *elaboración propia.*

3.4. Sistema Operativo

Ejecuta los comandos ordenados por la unidad de control y asociados a una combinación de teclas, a través de una interrupción al sistema operativo. La activación del comando sin que se presione físicamente una tecla, se logra con la librería *AutoIt* [40].

3.5 Síntesis de Voz

Siguiendo la instrucción de la unidad de control, reproduce un audio que informa del comando en ejecución. La reproducción del mensaje del audio (columna 4, Tabla 1) usa la librería *System.Speech.Synthesis* [41].

4. PRUEBAS Y RESULTADOS

Se realizaron 3 fases de pruebas. La primera busca establecer la experiencia previa del usuario en el manejo de cada aplicación. La segunda evalúa la operación del sistema. En estas primeras dos pruebas participaron personas sin limitaciones motrices (grupo 1 y 2), y son pruebas guiadas. La tercera fase evalúa el desempeño del sistema en personas con limitaciones motrices (grupo 3), usando el mismo protocolo guiada. Luego, se realizaron encuestas a los tres grupos de participantes para conocer su percepción del desempeño del sistema.

El grupo 1 estuvo compuesto por 8 hombres y 5 mujeres sin limitaciones motrices, con edades entre 22 y 32 años, familiarizados con el uso de computadores y navegación en internet. El grupo 2 estuvo compuesto por 7 hombres y 6 mujeres sin limitaciones motrices, con edades entre 45 y 73 años, y con poca experiencia en el uso de computadores. El grupo 3 estuvo compuesto por 4 hombres y 3 mujeres, con edades entre 33 y 57 años, con limitaciones motrices debidas a un trauma raquimedular, y con experiencia en el uso de computadores y navegación en internet.

4.1. Pruebas Fase 1

Se realizaron encuestas a las personas de los grupos 1 y 2 para establecer la experiencia previa en el manejo de computadores, navegación en internet, comandos de voz e interacción con cada una de las aplicaciones. Las respuestas a las preguntas previas se resumen en la Tabla 2 y la Fig. 3.

Tabla 2: *Encuesta Fase 1.*

	0-2 Horas	2-4 Horas	4-6 Horas	6-8 Horas	> 8 Horas
¿Cuántas horas al día usa el computador?					
Grupo 1	0%	0%	0%	46%	54%
Grupo 2	47%	30%	15%	8%	0%
¿Desde cuándo usa Internet?					
	6-12 meses	1-2 años	2-5 años	> 5 años	
Grupo 1	0%	0%	0%	100%	
Grupo 2	24%	38%	23%	15%	
¿Cuántas horas diarias navega en internet?					
	0-2 Hours	2-4 Hours	4-6 Hours	6-8 Hours	> 8 Hours
Grupo 1	0%	0%	23%	30%	47%
Grupo 2	54%	38%	8%	0%	0%
¿Ha usado sistemas de reconocimiento de voz?					
	Usually	Regularly	Rarely	Never	
Grupo 1	8%	15%	30%	47%	
Grupo 2	0%	8%	23%	69%	

Fuente: *elaboración propia.*

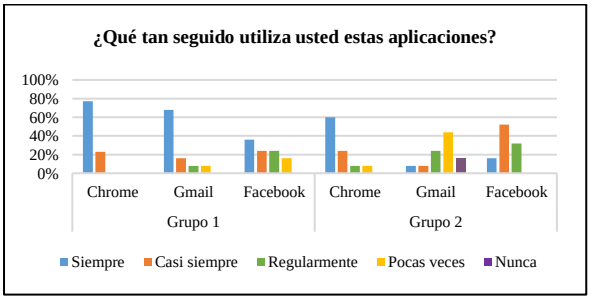

Fig. 3. *Respuestas a la pregunta ¿Qué tan a menudo utiliza usted estas aplicaciones?*
Fuente: *elaboración propia.*

Se observa que los miembros del grupo 1 usan el computador más de 6 horas al día, y han usado internet por más de 5 años. El 77 % de los miembros del grupo 2 usan el computador 4 horas diarias o

menos, y aunque el tiempo que llevan utilizando Internet es muy variado, la mayoría (62%) lo hace desde hace menos de dos años.

En orden decreciente de porcentajes, el grupo 1 usa Google, Gmail y Facebook, mientras que el grupo 2 usa Google, Facebook y Gmail. Ambos grupos han tenido poca o ninguna experiencia con sistemas de reconocimiento del habla, aunque el grupo 1 está ligeramente más familiarizado con su uso.

Los resultados obtenidos sugieren que las interfaces con mecanismos de interacción hombre-máquina, tales como comandos de voz, podrían facilitar el uso de aplicaciones populares cuando no se cuenta o hay muy poca experiencia con redes sociales (tales como el grupo 2).

4.2. Pruebas Fase 2

Se realizaron pruebas con los grupos 1 y 2, utilizando diálogos guiados que incluyeron comandos de voz y dictados en cada una de las tres aplicaciones. Las pruebas evaluaron la capacidad de reconocimiento de los comandos y las palabras que conforman el dictado. Se realizó una encuesta al finalizar las pruebas, para determinar la percepción de los usuarios e identificar aspectos a mejorar en el sistema de interacción por voz.

Las aplicaciones se probaron en tres sesiones espaciadas por una semana, probando una por semana. Cada repetición requirió de tres minutos aproximadamente, mientras que los dictados toman menos tiempo que escribir el texto por teclado. Las pruebas se describen a continuación, discutiendo los resultados para cada aplicación.

4.2.1. Prueba de la Aplicación Google Chrome

En esta prueba se pidió a los usuarios abrir la aplicación mediante el comando "***Abrir Google***", iniciar una búsqueda, y pronunciar la frase "***El destino es el que baraja las cartas y nosotros somos los que jugamos***". Una vez visualizados los resultados, el usuario debía pronunciar la palabra "***Navegar***" para desplazarse secuencialmente sobre ellos, utilizando los comandos "***Siguiente***" o "***Anterior***". Cuando el usuario selecciona el resultado que desea visitar, debía pronunciar el comando "***Entrar***". Posteriormente debía decir el comando "***Acercar***" que activa el zoom in del 25% y, finalmente, el comando "***Bajar***" para desplazarse verticalmente sobre el documento. A continuación, debía pronunciar "***Alejar***" para activar el zoom out del 25%, seguido del comando "***Subir***" para desplazarse sobre el documento y, "***Actualizar***" para refrescar la información desplegada en el navegador. Para acceder a otra búsqueda, el usuario debía decir "***Atrás***" y, finalmente, decir el comando "***Nueva búsqueda***" usando la frase de búsqueda "***El destino es el puente que construyes hacia lo que quieres***".

El desempeño del sistema se cuantificó por el porcentaje promedio de comandos y palabras de dictado identificados correctamente [42] y se ilustra en la Fig. 4.

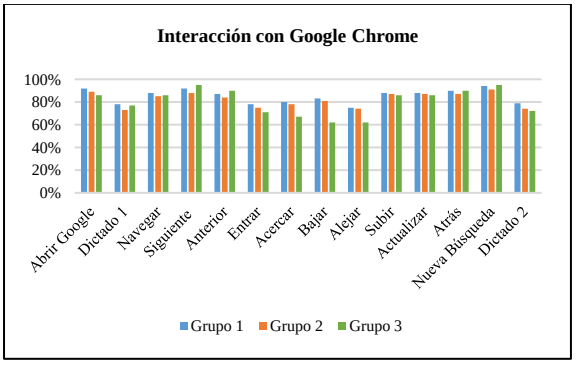

Fig. 4. *Porcentajes de desempeño del reconocimiento vocal para Google Chrome.*
Fuente: *elaboración propia.*

Los comandos con mejor desempeño fueron "***Nueva Búsqueda***", "***Abrir Google***" y "***Siguiente***" con un 92.83%, 90.83% y 90.50%, respectivamente, mientras que los comandos con menor desempeño fueron "***Alejar***", "***Entrar***" y "***Acercar***", con promedios del 74.83%, 76.50% y 79.50%, respectivamente. En estas palabras, se presentan características fonéticas similares como el que todas terminan en "***ar***". Los inconvenientes para la correcta identificación de los comandos terminados en "***ar***" pueden deberse a la pronunciación del fonema "***r***", que fue detectada en aproximadamente el 23.07% de los usuarios.

El sistema de reconocimiento de comandos presentó un desempeño ligeramente superior para el grupo 1 (promedio del 86.64% y desviación estándar del 6.04%) frente al grupo 2 (promedio 84.08% y desviación estándar del 5.57%). Lo mismo ocurrió en el reconocimiento del dictado: el grupo 1 alcanzó un promedio 78.50% con desviación estándar de 5.54%, frente a un promedio 74% y desviación estándar de 3.22% del grupo 2.

La correcta identificación de comandos de voz alcanzó un promedio conjunto de 85.36% con una desviación estándar del 5.77%, mientras que el

dictado alcanzó un 76.25% con una desviación estándar del 4.92%.

Después de la interacción de los usuarios con la aplicación Google Chrome, se les realizó una encuesta cuyas respuestas se presentan en las Fig. 5 y 6.

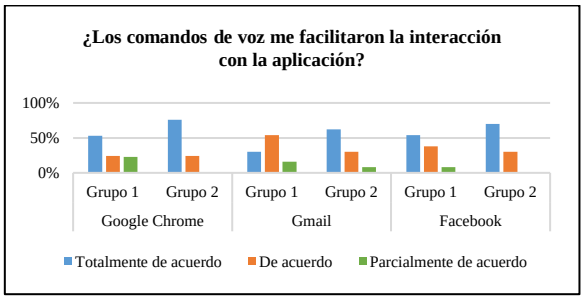

Fig. 5. *Resultados cualitativos en la interacción con las aplicaciones.*
Fuente: *elaboración propia.*

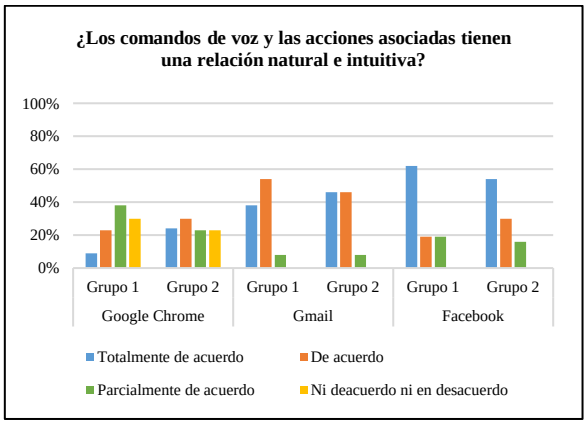

Fig. 6. *Resultados cualitativos de los comandos para las aplicaciones.*
Fuente: *elaboración propia.*

La mayoría de los usuarios de ambos grupos encontró útil la interacción con comandos de voz; no obstante, el grupo 2 lo encontró más útil, lo que se reflejada en una diferencia del 23%. El 23% del grupo 1 estuvo parcialmente de acuerdo con la utilidad de los comandos de voz, posiblemente por una mayor familiarización con las formas tradicionales de interacción.

Más del 40% de los usuarios fueron neutrales o parcialmente de acuerdo respecto a la relación natural e intuitiva de los comandos y las acciones asociadas; además, recomendaron usar verbos conjugados, como por ejemplo "***Abre Google***" en lugar de "***Abrir Google***". Igualmente, sugirieron cambiar el comando "***Entrar***" por "***Abrir***", ya que se refiere naturalmente a "***Abrir Enlace***".

4.2.2. Prueba de la Aplicación Gmail

La meta de esta prueba era enviar un correo electrónico al destinatario "Hernando", e introducir un asunto y mensaje preestablecidos. La secuencia de comandos de voz y dictados se presentan secuencialmente en la Tabla 3.

Tabla 3: *Protocolo de pruebas para Gmail.*

Tipo de interacción	Comando de Voz
Comando	Abrir Gmail
	Nuevo correo
Dictado 1	"Hernando"
	Siguiente
Comando	Aceptar
	Agregar Asunto
Dictado 2	"Recordatorio segunda reunión"
Comando	Agregar mensaje
Dictado 3	"Estimado Hernando, por medio de la presente le recuerdo que la próxima reunión del grupo de investigación se llevará a cabo el día lunes 30 de febrero del año en curso"
	Enviar Correo
	Correos Enviados
Comando	Entrar
	Abrir Mensaje

Fuente: *elaboración propia.*

El desempeño del sistema se cuantificó [42] como el porcentaje promedio de comandos y palabras de dictado identificados correctamente, y se ilustra en la Fig. 7.

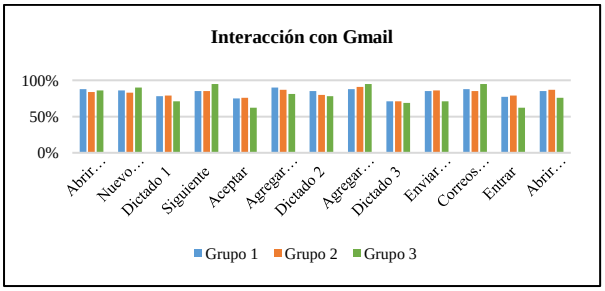

Fig. 7. *Porcentajes de desempeño en el reconocimiento vocal para Gmail.*
Fuente: *elaboración propia.*

Los resultados son consistentes con los obtenidos para la aplicación Google Chrome, con un 77.33% de desempeño y una desviación estándar de 5.84% para los dictados, y un 84.5% con una desviación estándar de 4.4% para los comandos. Los comandos "***Aceptar***" y "***Entrar***" fueron los que presentaron un desempeño inferior al 80%; mientras que el comando "***Siguiente***" y los comandos compuestos por dos palabras fueron los que presentaron un desempeño mayor al 85% para ambos grupos.

El bajo desempeño para el dictado 1 (78.83%) se debió a que está compuesto por una sola palabra:

"Hernando", que además puede confundirse con palabras que poseen características sonoras similares, tal como "Fernando".

El dictado del asunto (dictado 2) compuesto por tres palabras tuvo un desempeño promedio del 83%, mientras que el dictado del mensaje (dictado 3) constituido por 31 palabras, tuvo un desempeño promedio del 71.50%. Lo anterior confirma la limitación en el reconocimiento de voz para dictados largos, pues hace más complejo y menos preciso el análisis sintáctico involucrado.

En la interacción con Gmail (Fig. 5), la mayoría de usuarios de ambos grupos encontró útil la interacción con comandos de voz. Sin embargo, el grupo 2 revela una mayor aceptación reflejada en un porcentaje de 92% frente al 84% del grupo 1. No obstante, el 16% del grupo 1 y el 8% del grupo 2 están parcialmente de acuerdo con el uso de estos comandos.

En los comandos para Gmail (Fig. 6), más del 90% de los usuarios consideraron que los comandos fueron intuitivos, pero hicieron recomendaciones similares a las de la prueba con Google Chrome sobre el uso de verbos conjugados.

4.2.3. Prueba de la Aplicación Facebook

El objetivo de esta prueba es evaluar la aplicación de Facebook, navegando por las secciones Notificaciones, Mensajes, Perfil y Muro, y finalmente, modificar su estado. La secuencia de comandos de voz y dictados se consignan secuencialmente en la Tabla 4.

Tabla 4: *Protocolo de pruebas en Facebook.*

Tipo de interacción	Comando de Voz
Comando	Abrir Facebook Abrir notificaciones Abrir mensajes Abrir perfil Abrir muro Nuevo estado
Dictado	"El hombre fuerte es el que es capaz de interceptar a voluntad la comunicación entre los sentidos y la mente."
Comando	Publicar estado

Fuente: *elaboración propia.*

El desempeño del sistema se cuantificó [42] como el porcentaje promedio de comandos y palabras de dictado identificados correctamente, y se ilustra en la Fig. 8. Los comandos con mejor desempeño fueron "***Abrir Facebook***", "***Publicar Estado***" y "***Abrir Notificaciones***" con un 92.67%, 92.17% y 91.67%, respectivamente; mientras que los comandos con el menor desempeño fueron "***Abrir Perfil***" y "***Abrir Muro***" con promedios del 84.00% y 84.50%, respectivamente. En este caso los porcentajes de desempeño fueron mayores que con las aplicaciones de Gmail y Google Chrome, debido a que todos los comandos involucraban dos palabras.

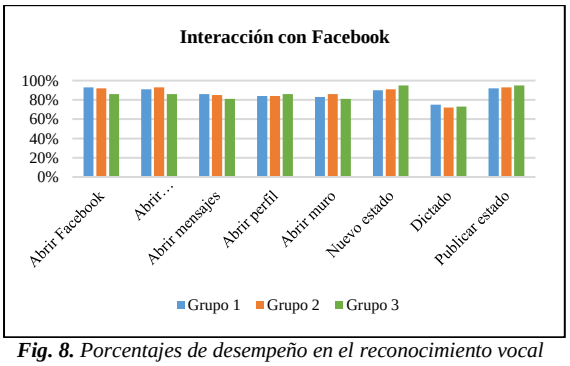

Fig. 8. *Porcentajes de desempeño en el reconocimiento vocal para Facebook.*
Fuente: *elaboración propia.*

Los resultados del reconocimiento de comandos presentan poca variabilidad entre los dos grupos, con un porcentaje global del 88.74% y una desviación estándar del 3.85%; mientras que en el dictado el desempeño fue menor (73.5%, con desviación estándar de 3.62%). Los resultados del dictado en las aplicaciones parecen tener una relación inversamente proporcional con el número de palabras (20 en este caso) y 11 palabras en la prueba de Google Chrome (palabras).

En la interacción con Facebook (Fig. 5), la mayoría de usuarios de ambos grupos encuentran útil la interacción con comandos de voz, esta vez con un promedio global mayor a 96%. El grupo 2 aceptó la utilidad en un 100%.

En los comandos para Facebook (Fig. 6), la mayoría de los usuarios de los grupos 1 y 2 (82.5%) sintieron que la interfaz es natural e intuitiva de la interfaz. Sin embargo, sugirieron omitir la palabra "***Abrir***" presente en la mayoría de los comandos, para brindar una interacción más rápida.

4.2.4. Comparación del desempeño del sistema

El reconocimiento de comandos alcanzó un rendimiento del 85,36% para Google Chrome, del 84,17% para Gmail y del 88,74% para Facebook. El rendimiento global fue del 86,09%, con una desviación estándar del 3,13%. El reconocimiento del dictado alcanzó un rendimiento medio global del 75,67%, con una desviación estándar del 5,11%.

Estos resultados indican que el reconocimiento de comandos y dictados entre aplicaciones es similar.

Se observa una variación en el desempeño con una relación inversa entre el reconocimiento y el número de palabras dictadas, lo que corresponde a una limitación del sistema propuesto. Una vez concluidas las pruebas en las tres aplicaciones descritas, se realizó una nueva encuesta para caracterizar la percepción de los usuarios de los grupos 1 y 2 sobre el rendimiento general de la interfaz desarrollada. Las preguntas fueron las siguientes:

Pregunta 1. ¿La herramienta me permitió una interacción amigable con el computador?
Pregunta 2. ¿Los comandos de voz fueron identificados correctamente por el sistema?
Pregunta 3. ¿El dictado fue reconocido correctamente por el sistema?
Pregunta 4. ¿En la medida en que interactué con el sistema pude activar los comandos con mayor facilidad?
Pregunta 5. ¿La realimentación auditiva contribuyó a tener una mejor experiencia?

Los resultados para los grupos 1 y 2 se presentan en las Fig. 9 y 10, respectivamente. La mayoría de los usuarios de ambos grupos califican la interfaz como amigable, con un promedio global de 92.3% entre "totalmente de acuerdo" y "de acuerdo".

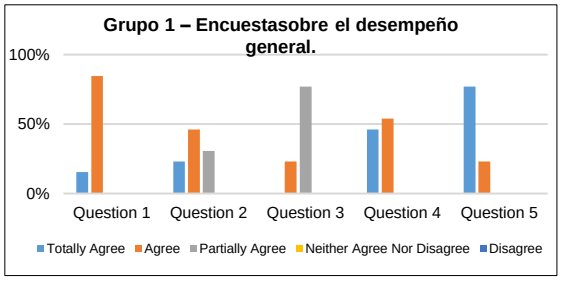

Fig. 9. *Grupo 1 – Encuesta del desempeño general del sistema.* ***Fuente:*** *elaboración propia.*

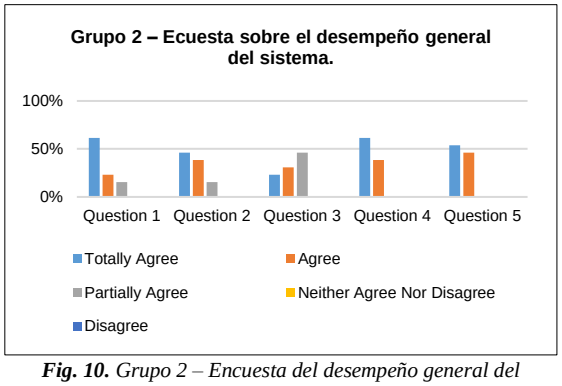

Fig. 10. *Grupo 2 – Encuesta del desempeño general del sistema.* ***Fuente:*** *elaboración propia.*

Los comandos alcanzaron un promedio global de 76.92% y los dictados 38.46%, con respuestas de totalmente de acuerdo" o "de acuerdo. Las respuestas sobre la percepción del reconocimiento de los comandos y de dictados varían en cada grupo, aunque hay una mayor satisfacción con el reconocimiento de comandos que a los dictados.

Los grupos 1 y 2 conceptuaron que, a medida que interactuaban con el sistema podían activar los comandos más fácilmente (100%), si se suman las respuestas de "totalmente de acuerdo" (53.85%) y las "de acuerdo" (46.15%).

Finalmente, en la encuesta se incluyó un campo abierto para que los usuarios aportaran sugerencias sobre cómo mejorar la interacción con la interfaz. Entre las principales recomendaciones estaban reforzar el reconocimiento del dictado, poder editar el texto dictado (seleccionar, copiar y pegar) y la inclusión de signos de puntuación.

4.3. Pruebas Fase 3

En la fase 3 participó un grupo de adultos con edades entre 33 y 57 años. Estas personas tenían lesiones a nivel cervical entre C3 y C7, pero sin problemas a nivel cognitivo o de pronunciación. Estas personas son los usuarios ideales para controlar el sistema de cómputo y realizar tareas básicas de navegación por sí mismos.

Al inicio de la prueba, se realizaron ejercicios de familiarización con las aplicaciones utilizando mouse y teclado, para luego presentarles la interfaz comandada por voz para la aplicación específica. Como en las pruebas anteriores, el desempeño del sistema se cuantificó por el porcentaje promedio de los comandos y dictados correctamente identificados [42].

La Fig. 4 muestra los resultados obtenidos para la aplicación Google Chrome. El desempeño promedio fue de 81.35% para comandos y 74.75% para dictados, con desviaciones estándar de 12.43% y 3.48%, respectivamente. Estos resultados son inferiores en alrededor de un 4% y 2% respecto a los obtenidos por los grupos 1 y 2 (85.36% y 76.25%), siendo los comandos "***Bajar***" y "***Alejar***" con los que se obtuvo el menor desempeño de los comandos

La Fig. 7 presenta los resultados para Gmail, con un promedio de 81.43% para comandos y 72.67% para dictados, y desviaciones estándar de 13.18% y 4.61%, respectivamente. Como puede observarse, el desempeño disminuyó en aproximadamente un 3%

para comandos y 5% para dictados respecto a los resultados obtenidos por los grupos 1 y 2 (84.17% y 77.25%), siendo "***Aceptar***" y "***Entrar***" los comandos con menor reconocimiento.

La Fig. 8 presenta los resultados obtenidos para la aplicación de Facebook, con un 87,07% de éxito en el reconocimiento para comandos y 73.33% para los dictados, y desviaciones estándar de 5,97% y 2.58% respectivamente, con una diferencia aproximada del 1% para comandos y dictados respecto a los obtenidos por los grupos 1 y 2 (88.74% y 73,50%). Todos los comandos presentaron un desempeño que varía entre 80.95% y 85.71% a excepción de los comandos "***Nuevo Estado***" y "***Publicar Estado***" (95.24%).

Los resultados globales de las pruebas a personas con limitaciones motrices arrojan una tasa de reconocimiento del 83.28%, con una desviación del 10.52% para comandos, y de 73.59% con una desviación del 3.56% para dictados. Estos resultados son inferiores aproximadamente en un 3% respecto a los resultados generales en los grupos 1 y 2 (86.09% y 75.67%). La reducción se atribuye a fatiga en la voz observada en las personas con limitaciones motrices y a afectaciones vocales. Un decremento del 3% es aceptable y es un indicador de robustez de la interfaz de comandos de voz bajo diferentes perfiles de usuario.

Después que se completó la prueba, se realizó la encuesta descrita arriba para caracterizar la percepción de los usuarios. En los resultados de la Fig. 11, la mayoría de los usuarios eligió la opción "*de acuerdo*", evidenciando una interacción amigable con las aplicaciones, así como que la realimentación auditiva contribuyó a mejorar la experiencia.

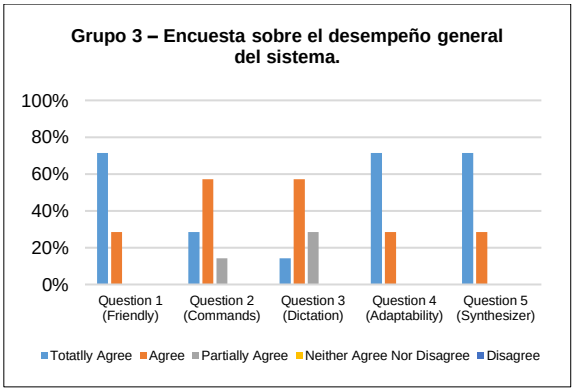

Fig. 11. *Grupo 3 – Encuesta cobre el desempeño general del sistema.*
Fuente: *elaboración propia.*

5. CONCLUSIONES

Se desarrolló una interfaz para controlar las aplicaciones Google Chrome, Gmail y Facebook mediante comandos de voz, usando las librerías de *Bing Speech API* de *Microsoft Cognitive Services* para el reconocimiento del habla, y un sintetizador de voz a través de la librería *System Speech Synthesis* para realimentar auditivamente al usuario. De esta manera, se ofrece una interacción humano-máquina más natural, y con alto potencial para ser utilizado por personas con limitaciones visuales o motrices que, de otro modo, no podrían interactuar con las interfaces tradicionales. El sistema propuesto fue diseñado para adaptarse fácilmente a otras aplicaciones en internet, tales como YouTube, Instagram y Twitter, y para controlar aplicaciones multimedia y de oficina.

Las pruebas se realizaron con tres grupos de personas de diferentes edades y niveles de familiarización con las aplicaciones. Uno de los grupos (grupo 3) estuvo conformado por personas con limitaciones motrices en sus miembros superiores e inferiores, algunas con afectaciones en las cuerdas vocales, una condición desafiante para el sistema. Los resultados de los grupos 1 (jóvenes) y 2 (adultos) arrojaron un promedio general en el reconocimiento de comandos de voz del 86.09% con una desviación estándar de 3.13%; y un promedio de 75.67% con una desviación estándar del 5.11% para el reconocimiento de dictados. El tercer grupo alcanzó un promedio general del 83.28% con una desviación del 10.52% en el reconocimiento de comandos, y un 73.63% con una desviación del 2.75% en el reconocimiento de dictados.

Los comandos de voz compuestos por dos o tres palabras fueron mejor identificados que aquellos de una sola palabra, como "***Aceptar***", "***Entrar***", "***Alejar***" y "***Acercar***" con los que se obtuvieron los menores porcentajes, debido a que hay un alto número de palabras con características fonéticas similares dentro del diccionario Español-Mexicano usado, y a problemas de dicción del fonema "***r***" identificado en aproximadamente el 23.07% de los usuarios. En los dictados se observó una relación inversa entre el desempeño promedio y el número de palabras que los componen, demostrando la dificultad de procesar secuencias de audios largos por los sistemas de reconocimiento del habla.

Los usuarios de los grupos 2 y 3 acogieron favorablemente la interacción verbal, mientras que los del grupo 1 mostraron una tendencia a preferir el uso de las interfaces tradicionales, seguramente por

estar adaptados a ellas. Los usuarios con limitaciones motrices manifestaron que la interfaz tiene un impacto positivo en su independencia informática y que, al igual que el resto de usuarios, en la medida que se familiarizaron con ella mejoraron su manejo. La gran mayoría de usuarios sintió que la realimentación auditiva mejoró su experiencia en el uso de la interfaz.

La integración de un sistema de identificación de usuario por voz permitiría una interfaz personalizable y totalmente independiente para cada usuario, además, se pueden incluir modelos de pronunciación de cada usuario para mejorar el rendimiento del reconocimiento de voz, dejando la posibilidad de integrar y eliminar comandos según el uso de cada usuario.

El sistema presentó pequeñas variaciones significativas para los diferentes grupos de estudio en los porcentajes de desempeño del reconocimiento de comandos de voz y dictado. Las variaciones en los porcentajes se debieron principalmente a la dicción y pronunciación de los usuarios.

Por último, se incluyó en la encuesta un campo abierto para que los usuarios aportaran sugerencias sobre cómo mejorar la interacción con la interfaz. Entre las principales recomendaciones estaban reforzar el reconocimiento del dictado, poder editar el texto dictado (seleccionar, copiar y pegar) y la inclusión de signos de puntuación.

REFERENCIAS

[1] N. S. Sreekanth et al., “Multimodal interface for Effective Man Machine Interaction,” Media Convergence Handbook, Media Business and Innovation, vol. 2, pp. 249–264, 2016. doi: 10.1007/978-3-642-54487-3.

[2] B. Basharirad and M. Moradhaseli, “Speech Emotion Recognition Methods: A Literature Review,” 2nd Int. Conf. Appl. Sci. Technol. 2017, 2017, doi: 10.1063/1.5005438.

[3] H. Ibrahim and A. Varol, “A Study on Automatic Speech Recognition Systems,” 8th Int. Symp. Digit. Forensics Secur. ISDFS 2020, p. 1–5, https://doi.org/10.1109/ISDFS49300.2020.9116 2, 2020.

[4] S. Raju, V. Jagtap, P. Kulkarni, M. Ravikanth, and M. Rafeeq, “Speech Recognition to Build Context: A Survey,” Int. Conf. Comput. Sci. Eng. Appl., p. 1–7, https://doi.org/10.1109/iccsea49143.2020.9132, 2020.

[5] L. Oksana, T. Ihor, and L. Pavlo, “Navigation Assistive Application for the Visually Impaired People,” Proc. - 2020 IEEE 11th Int. Conf. Dependable Syst. Serv. Technol. DESSERT 2020, p. 320–325, https://doi.org/10.1109/DESSERT50317.2020, 2020.

[6] L. Clark et al., “The State of Speech in HCI: Trends, Themes and Challenges,” Interact. Comput., vol. 31, no. 4, p. 349–371, https://doi.org/10.1093/iwc/iwz016, 2019.

[7] Apple, “Siri,” [Online - Accesed May 4, 2023]. [Online]. Available: http://www.apple.com/siri/

[8] Microsoft, “Cortana,” [Online - Accesed May 4, 2023]. [Online]. Available: https://www.microsoft.com/en-us/cortana

[9] Google, “Google Assistant,” [Online - Accesed May 4, 2023]. [Online]. Available: https://assistant.google.com/intl/es_es/

[10] Amazon, “Alexa,” [Online - Accesed May 4, 2023]. [Online]. Available: https://developer.amazon.com/alexa

[11] Samsung, “Bixby,” [Online - Accesed May 4, 2023]. [Online]. Available: https://www.samsung.com/us/explore/bixby/

[12] Cerence, “Cerence Drive,” [Online - Accesed May 4, 2023]. [Online]. Available: https://www.cerence.com/cerence-products/beyond-voice

[13] Amazon, “Astro,” [Online - Accesed May 4, 2023]. [Online]. Available: https://www.amazon.com/-/es/Presentamos-Amazon-Astro/dp/B078NSDFSB

[14] R. Sarikaya, “The Technology Behind Personal Digital Assistants,” IEEE Signal Process. Mag., vol. 34, p. 67–81, https://doi.org/10.1109/MSP.2016.2617341, 2017.

[15] V. Kepuska and G. Bohouta, “Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home),” 2018 IEEE 8th Annu. Comput. Commun. Work. Conf. CCWC 2018, vol. 2018–Janua, no. c, pp. 99–103, 2018, doi: 10.1109/CCWC.2018.8301638.

[16] E. Marvin, “Digital Assistant for the Visually Impaired,” 2020 Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2020, p. 723–728, https://doi.org/10.1109/ICAIIC48513.2020., 2020.

[17] M. B. Chandu and K. Ganapathy, “Voice Controlled Human Assistence Robot,” 2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020, p. 971–973,

https://doi.org/10.1109/ICACCS48705.2020., 2020.

[18] S. Faroom, M. Nauman, S. Yousaf, and S. Umer, "Literature Review on Home Automation System," Int. Conf. Comput. Math. Eng. Technol., 2018, doi: 10.17148/IJARCCE.2017.63173.

[19] P. Suesaowaluk, "Home Automation System Based Mobile Application," 2nd World Symp. Artif. Intell., p. 97–102, https://doi.org/10.1109/wsai49636.2020.914, 2020.

[20] N. H. Abdallah, E. Affes, Y. Bouslimani, M. Ghribi, A. Kaddouri, and M. Ghariani, "Smart Assistant Robot for Smart Home Management," 1st Int. Conf. Commun. Control Syst. Signal Process., p. 317–321, https://doi.org/10.1109/ccssp49278.2020.9, 2020.

[21] P. J. Rani, J. Bakthakumar, B. P. Kumaar, U. P. Kumaar, and S. Kumar, "Voice controlled home automation system using natural language processing (NLP) and internet of things (IoT)," ICONSTEM 2017 - Proc. 3rd IEEE Int. Conf. Sci. Technol. Eng. Manag., vol. 2018–Janua, pp. 368–373, 2018, doi: 10.1109/ICONSTEM.2017.8261311.

[22] P. Dabre, R. Gonsalves, R. Chandvaniya, and A. V. Nimkar, "A Framework for System Interfacing of Voice User Interface for Personal Computers," 3rd Int. Conf. Commun. Syst. Comput. IT Appl., p. 1–6, https://doi.org/10.1109/cscita47329.2020.9137, 2020.

[23] V. Chayapathy, G. S. Anitha, and B. Sharath, "IOT based home automation by using personal assistant," Proc. 2017 Int. Conf. Smart Technol. Smart Nation, SmartTechCon 2017, pp. 385–389, 2018, doi: 10.1109/SmartTechCon.2017.8358401.

[24] L. P. De Oliveira, M. A. Wehrmeister, and A. S. De Oliveira, "Systematic Literature Review on Automotive Diagnostics," Brazilian Symp. Comput. Syst. Eng. SBESC, vol. 2017–Novem, pp. 1–8, 2017, doi: 10.1109/SBESC.2017.7.

[25] H. Zhang and C. Ye, "Human-Robot Interaction for Assisted Wayfinding of a Robotic Navigation Aid for the Blind," 12th Int. Conf. Hum. Syst. Interact. (HSI), Richmond, VA, USA, p. 137–142, https://doi.org/10.1109/HSI47298.2019.894, 2019.

[26] G. Lugano, "Virtual assistants and self-driving cars," Proc. 2017 15th Int. Conf. ITS Telecommun. ITST 2017, pp. 1–5, 2017, doi: 10.1109/ITST.2017.7972192.

[27] M. Kim, E. Seong, Y. Jwa, J. Lee, and S. Kim, "A Cascaded Multimodal Natural User Interface to Reduce Driver Distraction," IEEE Access, vol. 8, p. 112969–112984, https://doi.org/10.1109/ACCESS.2020, 2020.

[28] S. Estes, J. Helleberg, K. Long, M. Pollack, and M. Quezada, "Guidelines for speech interactions between pilot & cognitive assistant," ICNS 2018 - Integr. Commun. Navig. Surveill. Conf., pp. 1–23, 2018, doi: 10.1109/ICNSURV.2018.8384965.

[29] S. Nur, A. Mohamad, and K. Isa, "Assistive Robot for Speech Semantic Recognition System," 2018 7th Int. Conf. Comput. Commun. Eng., p. 50–55, ISBN: 9781538669921, 2018.

[30] M. A. Hossain, M. F. K. Khondakar, M. H. Sarowar, and M. J. U. Qureshi, "Design and Implementation of an Autonomous Wheelchair," 2019 4th Int. Conf. Electr. Inf. Commun. Technol. EICT 2019, p. 1–5, https://doi.org/10.1109/EICT48899.2019.906885, 2019.

[31] N. H. Khan, A. H. Arovi, H. Mahmud, K. Hasan, and H. A. Rubaiyeat, "Speech based text correction tool for the visually impaired," pp. 150–155, ISBN: 978-1-4673-9930–2, 2015.

[32] P. Bose, A. Malpthak, U. Bansal, and A. Harsola, "Digital assistant for the blind," 2017 2nd Int. Conf. Converg. Technol. I2CT 2017, vol. 2017–Janua, no. 2015, pp. 1250–1253, 2017, doi: 10.1109/I2CT.2017.8226327.

[33] K. Cofre, E. Molina, and G. Guerrero, "Voice controlled interface oriented memory loss assistance system for older adults," Iber. Conf. Inf. Syst. Technol. Cist., p. 24–27, https://doi.org/10.23919/CISTI49556.2020.91, 2020.

[34] J.-H. Mosquera-DeLaCruz, S.-E. Nope-Rodríguez, A.-D. Restrepo-Girón, and H. Loaiza-Correa, "Internet Access by Voice Commands (Source Code)," [Online - Accesed Jun 17, 2023]. Accessed: Jun. 17, 2023. [Online]. Available: https://github.com/nandostiwar/Internet_Access_by_Voice_Commands.git

[35] Logitech, "LogitechG430," [Online - Accesed May 4, 2023]. [Online]. Available: https://www.logitechg.com/es-roam/products.html?searchclick=gaming

[36] Microsoft, "Bing Speech API," [Online - Accesed May 4, 2023]. [Online]. Available: https://azure.microsoft.com/es-

es/services/cognitive-services/speech-to-text/#overview

[37] M. Assefi, M. Wittie, and A. Knight, “Impact of Network Performance on Cloud Speech Recognition,” 2015, doi: 10.1109/ICCCN.2015.7288417.

[38] A. Hannun et al., “Deep Speech: Scaling up end-to-end speech recognition,” Arxiv, pp. 1–12, 2014, doi: arXiv:1412.5567v2.

[39] J.-H. Mosquera-DeLaCruz, S.-E. Nope-Rodríguez, A.-D. Restrepo-Girón, and H. Loaiza-Correa, “Disability and Rehabilitation : Assistive Technology Human-computer multimodal interface to internet navigation,” Disabil. Rehabil. Assist. Technol., vol. 0, no. 0, p. 1–14, https://doi.org/10.1080/17483107.2020.1799444, 2020.

[40] C. AutoIt, “AutoIt Library,” [Online - Accesed May 4, 2023], 2023. [Online]. Available: https://www.autoitscript.com/site/autoit/

[41] Microsoft, “System Speech Synthesis,” [Online - Accesed May 4, 2023], 2023. [Online]. Available: https://msdn.microsoft.com/en-us/library/system.speech.synthesis.aspx

[42] M. Canelli, D. Grasso, and M. King, “Methods and Metrics for the Evaluation of Dictation Systems: a Case Study,” Proc. Second Int. Conf. Lang. Resour. Eval., p. 1–7, SematicScholar Corpus ID: 15527622, 2000.