ISSN: 1692-7257 - Volume 1 – Number 45 - 2025

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i45.2963

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

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

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

Received: *june 16, 2024.* ***Accepted:*** *december 20, 2024.* ***Published:*** *january 01, 2025.*

How to cite: *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, jan. 2025.*
Recovered from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/2963

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.

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.

1. INTRODUCTION

Since the creation of computers, the methods of interacting with applications (keyboard and mouse) were not designed considering the natural manner in which humans interact, unintentionally excluding people with upper limb motor impairment and even technologically illiterate people. In contrast, current human-machine interfaces seek a more natural interaction [1]. Given that speech is the fastest and most natural means of communication between humans [2], it constitutes an option for human-machine interaction, associating sequences of words with machine commands under challenging audio quality conditions, in which ambient noise, diction quality, accent, intonation, timbre, volume and other factors affect [3], [4]; however, it would allow overcoming the usage barriers imposed by traditional interfaces [5].

The presence of computers and the Internet in academic, work and personal environments has modified the forms of communication between humans [6]. The massive use of applications such as Google Chrome, Gmail, Facebook, among others, have permeated to a greater or lesser extent in these areas; in particular, in conditions of social confinement due to biosafety conditions.

Recognition and voice synthesis systems have been commercially integrated in mobile devices and virtual personal assistants. Highlighted among the first to be developed are SIRI [7], CORTANA [8], GOOGLE ASSISTANT [9], ALEXA [10] and BIXBY [11], from Apple, Microsoft, Google and Samsung, respectively. Other developments designed to help users with their daily tasks and provide easy access to structured data, web services and personal applications were consulted, including CERENCE DRIVE [12] for car assistance or ASTRO [13] to control home devices. These assistants require the combination of speech recognition technologies, natural language processing, dialogue management, language generation and text synthesis [14], [15], [16], [17].

Three trends in voice-commanded virtual personal assistants were distinguished. The first one oriented to home automation [18], [19], [20], [21], [22], [23]. The second oriented to vehicle navigation [24], [25], [26], [27], [28], [29], [30], which includes cars, wheelchairs, and airplanes. The third, within which the present work is framed, is oriented towards the access osf information and communication technologies (ICTs) for accessing information of daily interest such as voice dictation and text correction [31], weather forecast and news [32] and calendar management [33].

This research presents a system for Internet navigation through voice commands in Facebook, Gmail and Chrome, allowing the performance of dictations and the synthesis of voice signals to provide feedback to the user. The methodology, system description, tests and conclusions are presented below.

2. METHODOLOGY

The most socially recognized and adopted navigation applications were selected. Then, the most-used options associated with voice commands in each navigation application were identified. To identify the user's voice command, a recognition subsystem was integrated, and to make the status of the system known to the user, a voice synthesizer subsystem was introduced. Finally, the routines required for executing the actions of the commands associated with the applications were implemented via interruptions to the operating system.

The evaluation of the interface was carried out based on analyses of two sources of information: surveys that measure the users' level of satisfaction with the use of voice commands and dictations in the mentioned applications and performance metrics in the speech recognition.

3. SYSTEM DESCRIPTION

The system is composed of six blocks, as illustrated in Fig. 1. The process begins with the acquisition of the user´s voice signal through a microphone connected to the computer system. Immediately after, the speech recognition block processes the captured audio and converts it into text strings. The command validation block identifies whether the string content corresponds to one of the orders established in the command dictionary and sends this information to the control unit block, which defines the actions to execute through the operating system block; finally, the system provides auditory feedback to the user using the voice synthesis block. The source code of this developed system can be consulted in [34].

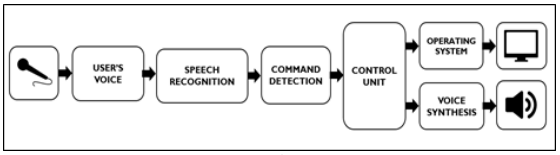

Fig. 1. *Diagram of the system blocks.*
Source: *own elaboration.*

3.1. Acquisition of the User's Voice

The acquisition of the user's voice was carried out with the built-in microphone in a Logitech G430 headset [35] in a semi-controlled environment with office-type noise conditions.

3.2. Speech Recognition

The proposed interface was developed in the C# programming language under the Windows 8.1 operating system. The speech recognition system is independent of the speaker and uses the Bing Speech API libraries [36] that belong to Microsoft Cognitive Services, ensuring the compatibility and integration with the rest of the algorithms. Use of these libraries requires a constant connection to the Internet, given that the speech recognition and its transcription is performed online [37]. The internal architecture of the speech recognition block is illustrated in Fig. 2 and consists of three models: Acoustic Model, Pronunciation Model and Language Model.

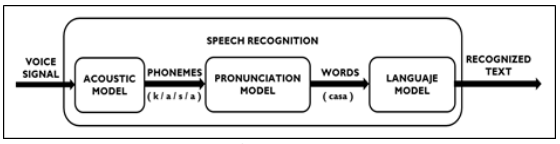

Fig. 2. *Block diagram of speech recognition.* ***Source:*** *own elaboration.*

The models that compose the speech recognition block are based on the probabilistic theory [38] for finding the word sequence that better fits the captured voice signal, considering the probability distributions for modeling the phonemes P(X|L), the pronunciation of phonemes within a defined dictionary P(L | W), and the language that organizes the words P(W), such that it is syntactically coherent, based on (1).

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

3.3. Command Validation

In this part of the system, a set of commands is compared with the recognized text of the user's voice, these commands are presented in Table 1 (second column) and are divided into two categories: generic commands, which are common to the three applications, and specific commands for each application. The same dictionary of commands proposed in [39] was used.

3.4. Control Unit

This unit receives the information coming from the command validation block and verifies if the command detected belongs to the application that is currently being executed. Then, it sends the corresponding orders to the operating system block and the voice synthesis block according to the lists in columns 3 and 4 of Table 1, respectively.

Table 1: *Commands, keyboard shortcuts and speech synthesis.*

	Command	Pressed Keys Interruption on the Operating System	Voice Synthesis Message
Generic Commands	"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{CTRLUP}")	"Este es el historial"
	"Imprimir"	Send("{CTRLDOWN}p{CTRLUP}")	"Configurando Impresión"
Gma	"Correos enviados"	Send("g"); Send("t")	"Abriendo mensajes enviados"

Command	Pressed Keys Interruption on the Operating System	Voice Synthesis Message
"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"

Source: *own elaboration.*

3.5. Operating System

It executes the commands ordered by the control unit and associated with a combination of keys through an interruption to the operating system. The commands activation, without needing to physically press the keys, is achieved with the AutoIt library [40].

3.6. Voice Synthesis

It provides an audio feedback to the user of the executed command following the instruction of the control unit. The playback of the audio messages (column 4, Table 1) uses the System.Speech.Synthesis library [41].

4. TESTS AND RESULTS

The tests were performed in three stages. The first seeks to establish prior user experience in managing each application, and in using voice commands. The second evaluates the system operation. These first two stages were performed with two groups of people without motor impairment (group 1 and group 2) and are guided tests. The third stage evaluates the system performance with a group of people with motor impairment (group 3), using the same guided-test protocol. Then, surveys were administered to the three groups of participants to learn their perception of the system performance.

Group 1 was composed of 8 men and 5 women without motor impairment, aged between 22 and 32 years old, who were familiar with computers and web browsing. Group 2 was composed of 7 men and 6 women without motor impairments, aged between 45 and 73 years old, and with little prior experience with computers. Group 3 was composed of 4 men and 3 women, aged between 33 and 57 years old, with motor impairment due to a spinal trauma and with experience in the use of computers and web browsing.

4.1. Stage 1 Tests

Surveys were conducted on the people of groups 1 and 2 to establish their prior experience with computers, Internet, voice commands and each application. The answers to the previous questions are summarized in Table 2 and Fig. 3.

Table 2: *Survey stage 1.*

¿How many hours per day do you use the computer?					
	0-2 Hours	2-4 Hours	4-6 Hours	6-8 Hours	> 8 Hours
Group 1	0%	0%	0%	46%	54%
Group 2	47%	30%	15%	8%	0%
¿How long have you been using the Internet?					
	6-12 Months	1-2 Years	2-5 Years		> 5 Years
Group 1	0%	0%	0%		100%
Group 2	24%	38%	23%		15%
¿How many hours per day do you browse the web?					
	0-2 Hours	2-4 Hours	4-6 Hours	6-8 Hours	> 8 Hours
Group 1	0%	0%	23%	30%	47%
Group 2	54%	38%	8%	0%	0%
¿Have you used speech recognition systems?					
	Usually	Regularly	Rarely	Never	
Group 1	8%	15%	30%	47%	
Group 2	0%	8%	23%	69%	

Source: *own elaboration.*

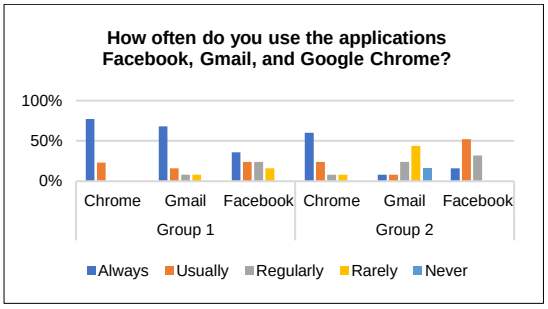

Fig. 3. *Answers to the question ¿How often do you use the applications Facebook, Gmail, and Google Chrome?.*
Source: *own elaboration.*

It was observed that the members of group 1 use the computer more than 6 hours per day, and have used the Internet for more than five years. 77% of the members of group 2 use the computer for 4 hours or less, and even though the length of time for which they have been using the Internet has a large range, most (62%) have been using it for less than two years.

In decreasing order of percentage, group 1 uses Google, Gmail and Facebook, whereas group 2 uses Google, Facebook and Gmail. Finally, both groups have had little or no use of speech recognition systems, although group 1 is slightly more familiar with their use.

The results obtained suggest that interfaces with human-machine interaction mechanisms, such as voice commands, could facilitate an easier use of these applications by populations with little or no experience with social networks (such as group 2).

4.2. Stage 2 Tests

Tests were performed with groups 1 and 2 using guided dialogs that included voice commands and dictations in each of the three applications. The tests evaluated the recognition capacity of both the commands as well as the words comprising the dictation. Furthermore, a survey was performed at the end of the tests to determine the users' perception and to identify aspects to improve upon in the voice interaction system.

The applications were tested in three sessions separated by one week, testing one application per week and performing each test three times. Each repetition required approximately three minutes, since dictation takes less time than writing. The tests are described below and the results are discussed for each application.

4.2.1. Google Chrome Application Test

In this test, the users were asked to open the application using the ***"Abrir Google" [Open Google]*** command, initiate the search, and say the phrase ***"El destino es el que baraja las cartas y nosotros somos los que jugamos" [Destiny is the one that shuffles the cards and we are the ones who play]***. Once the results were viewed, the user had to say the word ***"Navegar" [Browse]*** to sequentially move through the results using the commands ***"Siguiente" [Next]*** and ***"Anterior" [Previous]***. When the user selected the result he wanted to visit, he had to say the command ***"Entrar" [Enter]***. Then, the user was asked to say the word ***"Acercar" [Zoom-in]*** to activate the 25% zoom in and then say ***"Bajar" [Scroll down]*** to move along the scroll bar of the document. Subsequently, the user had to activate the 25% zoom out by saying ***"Alejar" [Zoom-out]***. Then, the user had to say ***"Subir" [Scroll up]*** to move through the document and ***"Actualizar" [Refresh]*** to refresh the information displayed by the browser. To access another search result, the user had to say ***"Atrás" [Back]*** and finally perform a new search by saying the command ***"Nueva búsqueda" [New search]*** and the search phrase ***"El destino es el puente que construyes hacia lo que quieres" [Destiny is the bridge that you build towards what you want]***.

The system performance was quantified by the average percentage of commands and dictated words correctly identified [42] and is illustrated in Fig. 4.

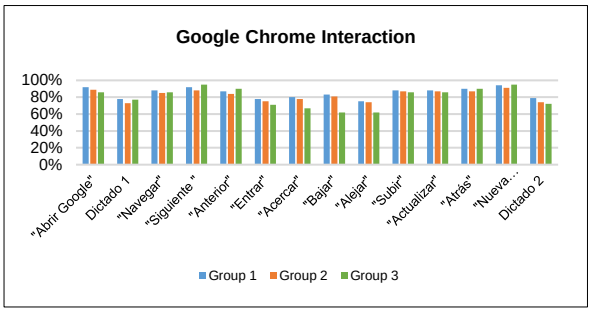

Fig. 4. *Speech recognition performance percentages for Google Chrome.*
Source: *own elaboration.*

The commands with the best performance were ***"Nueva Búsqueda" [New Search]***, ***"Abrir Google" [Open Google]*** and ***"Siguiente" [Next]***, with 92.83%, 90.83% and 90.50%, respectively, whereas the commands with lowest performance were ***"Alejar" [Zoom-out]***, ***"Entrar" [Enter]*** and ***"Acercar" [Zoom-in]***, with averages of 74.83%, 76.50% and 79.50%, respectively. In these words, there are similar phonetic characteristics, given that they all end in ***"ar"***. Hindering the accurate identification of commands ending in ***"ar"*** may be the pronunciation of the phoneme ***"r"***, which was detected in approximately 23.07% of users.

The command recognition system presented a slightly greater performance for group 1 (average of 86.64% and standard deviation of 6.04%) versus group 2 (average of 84.08% and standard deviation of 5.57%). The same behavior occurred in the dictation recognition: group 1 reached an average of 78.50% with a standard deviation of 5.54% versus an average of 74% and a standard deviation of 3.22% for group 2.

The correct identification of voice commands reached an overall average of 85.36% with a standard deviation of 5.77%, whereas the dictation reached 76.25% with a standard deviation of 4.92%.

After the interaction of the users with the Google Chrome application, a survey was carried out with

the questions whose answers are presented in Fig. 5 and 6.

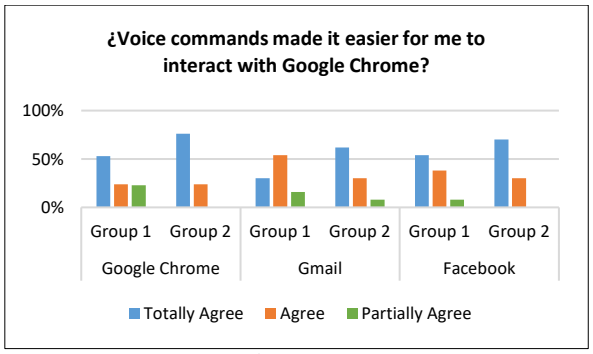

Fig. 5. *Qualitative results of interaction with Google Chrome.*
Source: *own elaboration.*

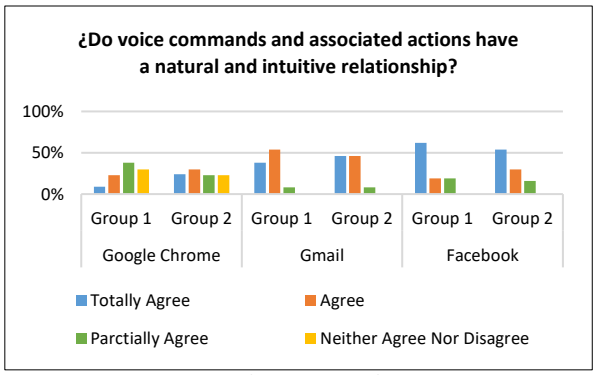

Fig. 6. *Qualitative results for commands for Google Chrome.*
Source: *own elaboration.*

Most users of both groups found the voice command interaction useful; however, group 2 found them more useful, as reflected by a difference of 23%. 23% of group 1 were in partially agreement about the usefulness of voice commands, possible related with their greater familiarization with traditional forms of interaction.

More than 40% of users were neutral or partially in agreement regarding the natural and intuitive relationship between the commands and associated actions; in addition, they recommend using conjugated verbs, such as ***"Abre Google" [Open Google]*** instead of ***"Abrir Google" [Open Google]***. In addition, it was suggested to change the command ***"Entrar" [Enter]*** to ***"Abrir" [Open]***, since ***'abrir'*** naturally refers to ***"Abrir Enlace" [Open Link]***.

4.2.2. Gmail Application Test

The goal of this test was to evaluate the Gmail application by sending an email to the recipient "Hernando", where the contents for the subject and the message were specified to the user. The sequence of voice commands and dictations is reported sequentially in Table 3.

Table 3: *Gmail tests protocol (in Spanish).*

Type of interaction	Voice Command
Command	Abrir Gmail Nuevo correo
Dictation 1	"Hernando"
Command	Siguiente Aceptar Agregar Asunto
Dictation 2	"Recordatorio segunda reunión"
Command	Agregar mensaje
Dictation 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"
Command	Enviar Correo Correos Enviados Entrar Abrir Mensaje

Source: *own elaboration.*

The system performance was quantified [42] as the average percentage of commands and dictated words correctly identified and is illustrated in Fig. 7.

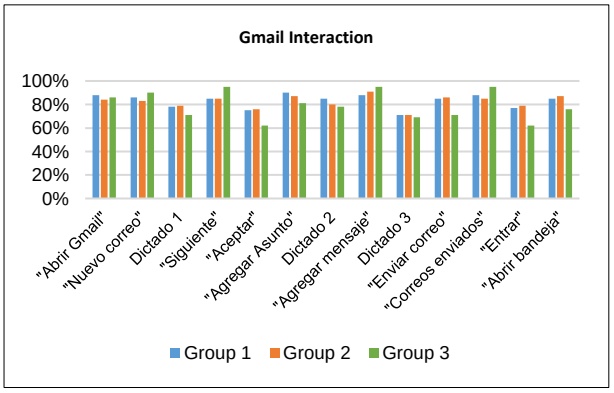

Fig. 7. *Speech recognition performance percentages for Gmail.*
Source: *own elaboration.*

The results are consistent with those obtained for the Google Chrome application, with a performance of 77.33% and a standard deviation of 5.84% for dictation and a performance of 84.5% with a standard deviation of 4.4% for commands. The commands "***Aceptar***" [Accept] and "***Entrar***" [Enter] were the ones that presented the lowest performance (80%), whereas the command "***Siguiente***" [Next] and the two-word commands presented a performance greater than 85% for both groups.

The low performance for dictation 1 (78.83%) was because it is composed of a single word, "Hernando", which can also be confused with words

that have similar sound characteristics, such as "Fernando".

The subject dictation (dictation 2), composed of three words, had an average performance of 83%, whereas the message dictation (dictation 3), composed of 31 words, had an average performance of 71.50% (the lowest obtained for the audio system). The above confirms the limitation in the speech recognition for long dictations, since it makes the syntactic analysis involved in the interpretation of a word set with meaning more complex and less accurate.

In the interaction with Gmail (Fig. 5), most users of both groups found the interaction through voice commands useful. However, group 2 reveals a greater acceptance, reflected in a percentage of 92%, compared to 84% for group 1. Nevertheless, 16% of group 1 and 8% of group 2 partially agreed with the use of these commands.

In the commands for Gmail (Fig. 6), more than 90% of users considered that the commands used were intuitive but made recommendations regarding the use of conjugated verbs, similar to the ones of the test with Google Chrome.

4.2.3. Facebook Application Test

The goal of this test is to evaluate the Facebook application by browsing the Notifications, Messages, Profile and Wall sections and, finally, to change their status. The sequence of voice commands and dictation is reported sequentially in Table 4.

Table 4: *Facebook tests protocol (in Spanish).*

Type of interaction	Voice Command
Command	Abrir Facebook Abrir notificaciones Abrir mensajes Abrir perfil Abrir muro Nuevo estado
Dictation	"El hombre fuerte es el que es capaz de interceptar a voluntad la comunicación entre los sentidos y la mente."
Command	Publicar estado

Source: *own elaboration.*

The system performance was quantified [42] as the average percentage of commands and dictated words identified correctly and is illustrated in Fig. 8. The commands with the best performance were ***"Abrir Facebook" [Open Facebook]***, ***"Publicar Estado" [Post Status]*** and ***"Abrir Notificaciones" [Open Notifications]***, with 92.67%, 92.17% and 91.67%, respectively, whereas the commands with lowest performance were ***"Abrir Perfil" [Open Profile]*** and ***"Abrir Muro" [Open Wall]***, with averages of 84.00% and 84.50%, respectively. In this case, the percentages were higher than with the Gmail and Google Chrome applications, given that all commands consisted of two words.

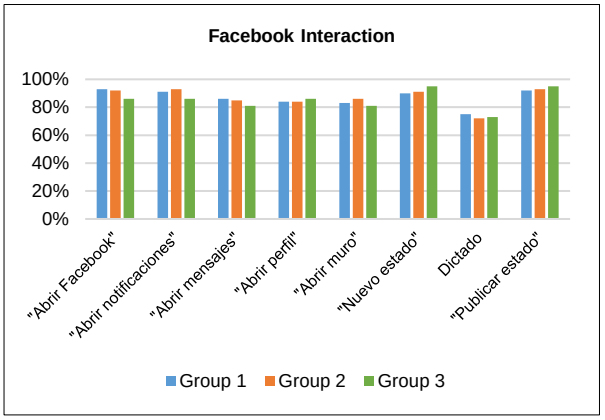

Fig. 8. *Performance percentages of speech recognition for Facebook.*
Source: *own elaboration.*

The command recognition performance varied little between the two groups, with an overall percentage of 88.74% and a standard deviation of 3.85%, whereas for dictation, a lower average performance was found (73.5%), with a standard deviation of 3.62%. Dictation thus exhibits an inversely proportional relationship between the number of words (in this case 20) and performance in the Google Chrome test (11 words).

In the interaction with Facebook (Fig. 5), most users of both groups find the interaction through voice commands useful, this time with an overall average of 96%. Group 2 exhibited 100% acceptance.

In the commands for Facebook (Fig. 6), most users of groups 1 and 2 (82.5%) felt that the interface is natural and intuitive. However, they suggested omitting the word "Abrir" [Open] for a faster interaction.

4.2.4 Comparison of the interface performance

The command recognition reached a performance of 85.36% for Google Chrome, 84.17% for Gmail, and 88.74% for Facebook. The overall performance was of 86.09%, with a standard deviation of 3.13%. The dictation recognition reached an overall average performance of 75.67% with a standard deviation of 5.11%. These results indicate similar command and dictation recognition between applications.

The variation of the recognition performance as a function of the number of dictated words, where a decreasing relationship is observed, associated with the speech recognition system's limitation against long strings of words.

After the tests on the three applications described above were completed, a new survey was conducted to characterize the perception of users of group 1 and 2 regarding the general performance of the developed interface. The questions were as follows:

Question 1. Did the tool give me a user-friendly interaction with the computer?
Question 2. Did the system correctly identify the voice commands?
Question 3. Did the system correctly identify the dictation?
Question 4. As I interacted with the system, was I able to activate the commands more easily?
Question 5. Did the audio feedback contribute to a better experience?

The results for groups 1 and 2 are presented in Fig. 9 and 10, respectively. Most users of both groups found the interface user-friendly, with an overall average of 92.3% between Fully agree and Agree.

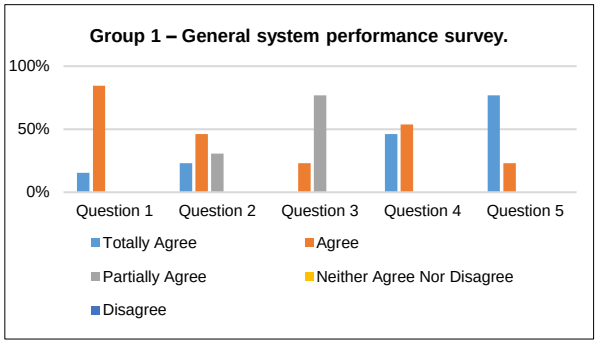

Fig. 9. *Group 1 – Survey for general system performance.*
Source: *own elaboration.*

Commands reached an overall average of 76.92%, and dictations 38.46%, for responses "fully agree" or "agree". The answers in response to user perception of command and dictation recognition vary in each group, although there was a greater satisfaction in both groups with command recognition compared to dictation.

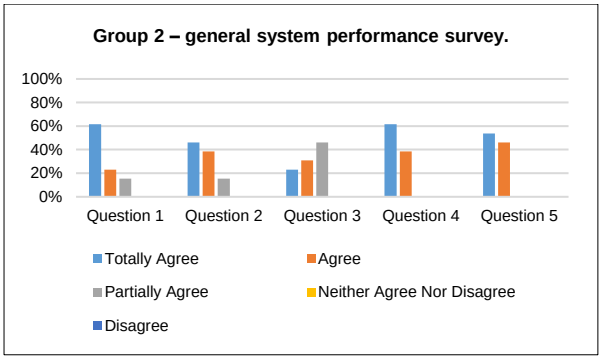

Fig. 10. *Group 2 - Survey for general system performance.*
Source: *own elaboration.*

Groups 1 and 2 felt as they interacted with the system that they could activate the commands more easily, reporting 100% between "fully agree" and "agree", distributed as 53.85% and 46.15%, respectively.

When evaluating the impact of the audio feedback in regard to the execution of commands, the users provided an overall response of 100% for "fully agree" (65.38%) and "agree" (34.62%), highlighting greater satisfaction among group 1 users.

Finally, an open field was included in the survey for users to provide suggestions for how to improve the interaction with the interface. Among the main recommendations were strengthening dictation recognition, being able to edit the dictated text (select, copy and paste), and the inclusion of punctuation marks.

4.3 Stage 3 Tests

During Stage 3, a group of adults between the ages of 33 and 57 was used. These people present a spinal cord lesion from C3 to C7, but do not present cognitive or pronunciation problems. These were ideal users of this interface, allowing them to control the computer system and perform basic web browsing tasks by themselves.

At the beginning of the tests, exercises were done to familiarize users with each of the applications (using the mouse and keyboard) to then present the voice command interface according to the specific application. As in the previous tests, the performance of the system was quantified as the average percentage of commands and dictated words correctly identified [42].

Fig. 4 shows the results obtained for the Google Chrome application. The average performance of the interface was of 81.35% for commands and 74.75% for dictations, with standard deviations of 12.43%

and 3.48%, respectively. These results are approximately 4% and 2% lower for commands and dictations with respect to those obtained in Stage 2 (85.36% and 76.25%), and the commands **"Bajar" [Down]** and **"Alejar" [Zoom-out]** remain the poorest performers.

Fig. 7 presents the results obtained for the Gmail application, with an average of 81.43% for commands and 72.67% for dictations, respectively, and standard deviations of 13.18% and 4.61%. As can be observed, the performance decreased by approximately 3% for commands and 5% for dictations with respect to the results obtained in Stage 2 (84.17% and 77.25%), with **"Aceptar" [Accept]** and **"Entrar" [Enter]** being the most difficult to recognize.

Fig. 8 presents the results obtained for the Facebook application, the average results indicate 87.07% success in command recognition and 73.33% for dictations, with standard deviations of 5.97% and 2.58%, respectively, representing a difference of approximately 1% for commands and dictations with respect to that obtained in Stage 2 (88.74% and 73.50%). The performance of all commands ranged between 80.95% and 85.71%, except for **"Nuevo Estado" [New Status]** and **"Publicar Estado" [Post Status]**, which had a performance of 95.24%.

The overall results of the tests in people with motor impairment reveal a recognition rate and standard deviation of 83.28% and 10.52% for commands, and 73.59% and 3.56% for dictation. These results are approximately 3% lower than the general results obtained for commands and dictations in Stage 2 (86.09% and 75.67%). This decrease may be attributed to the voice fatigue in the group with motor and speech impairment. An overall decrease of 3% is acceptable and is a good robustness indicator of the voice command interface under different user profiles.

After the tests were completed, the survey described above was conducted to characterize the users' perception. The results of Fig. 11, in which most users selected the option ***"De acuerdo" [Agree]***, show that the interface allows for a user-friendly interaction with the applications and that the audio feedback contributes to generating a better experience.

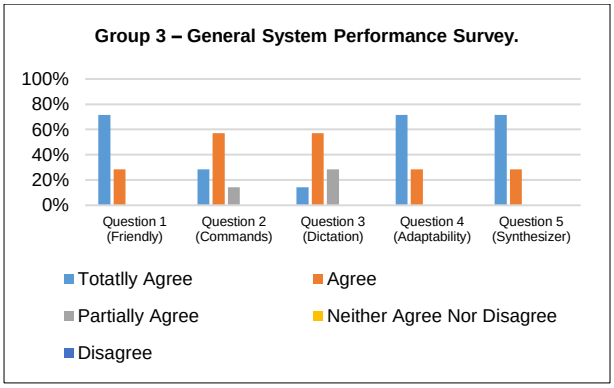

Fig. 11. *Group 3 – Survey of the general performance of the system.*
Source: *own elaboration.*

5. CONCLUSIONS

An interface was developed to control the applications Google Chrome, Gmail and Facebook through voice commands, using the *Bing Speech API libraries* of *Microsoft Cognitive Services* in speech recognition, and a voice synthesizer through the *System Speech Synthesis* library to provide the user with an audio feedback. In this manner, a more natural human-machine interaction is offered, with great potential to be used by people with visual or motor impairments that hamper the use of traditional interfaces. The proposed interface was designed to easily adapt to other applications oriented towards the Internet, such as YouTube, Instagram and Twitter, and also to control multimedia and office applications.

The tests were performed with three groups of people of different ages and levels of familiarization with the applications. One of the groups (group 3) was composed of people with motor impairments in their upper and lower limbs, some of them with affectations in their vocal cords, a challenging condition for testing the system. The results for groups 1 (young) and 2 (adults) yield a general recognition average for voice commands of 86.09% with a standard deviation of 3.13% and for dictation recognition an average of 75.67% with a standard deviation of 5.11%. The third group reached a general average and standard deviation of 83.28% and 10.52%, respectively, in command recognition and 73.63% and 2.75% in dictation recognition.

Voice commands composed of two or three words were better identified than the ones composed of one word. Commands such as "Aceptar" [Accept], "Entrar" [Enter], "Alejar" [Zoom-out] and "Acercar" [Zoom-in] obtained lower percentages due to the high number of words with similar phonetic characteristics contained in the Spanish-

Mexican dictionary used, and to the diction problems of the phoneme "**r**" in approximately 23.07% of users. An inverse relationship between the average performance and the number of words that compose each dictation was observed in the dictations, demonstrating the difficulty of processing long audio sequences presented by the speech recognition systems.

Users of groups 2 and 3 welcomed the vocal interaction, whereas users of group 1 tended to prefer traditional interfaces, surely because they are used to them. Users with motor disabilities stated that the interface had a positive impact in their using the computer independently, and similar to the rest of the users, as they became more familiar with it, managing the interface improved. Most users felt that the audio feedback through the voice synthesizer improved their experience during the use of the applications through the interface.

The system presented little significant variation in the recognition performance of voice commands and dictations for the different study groups. The variations in the percentages were mainly due to diction and pronunciation phenomena of the users.

The integration of a user identification system by voice would allow a customizable and totally independent interface for each user, in addition, pronunciation models for each user can be included to improve the performance of voice recognition, leaving the possibility of integrating and removing commands according to the use of each user.

REFERENCES

[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/iccsea49143.2020.913 2, 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.