Reconocimiento de comandos de voz en español orientado al control de una silla de ruedas

Spanish speech recognition oriented to a wheelchair control

L.J. Gil [1], L.F. Castillo[2], R.D. Flórez[3]

[1]Automática, Dpto. Electrón. y Automatización, UAM, Colombia, lily.gil@autonoma.edu.co;
[2] Docente TC Universidad de Caldas luis.castillo@udecaldas.edu.co - Docente Catedratico UAM lfcastil@autonoma.edu.co, Colombia;
[3] Automática, Dpto. de Electrónica y Automatización., UAM, Colombia, rubenfh@autonoma.edu.co

Recibido: nov 26, 2015. **Aceptado:** mar 3, 2016. **Versión Final:** mar 3, 2016

RESUMEN
Se presenta una aplicación computacional que reconoce instrucciones de voz en español para un vocabulario cerrado e independiente del hablante, adoptando el modelo de lenguaje que para el español proporciona la SAPI (Interfaz de Programación de Aplicaciones de Voz) de Microsoft®, de manera que reconozca solo la gramática relacionada con las funcionalidades que el usuario de la *silla de ruedas automatizada* que se trabaja al interior del grupo de investigación de Automática de la Universidad Autónoma de Manizales va a manejar. Las pruebas para medir el desempeño del sistema de reconocimiento se realizan de manera discriminada por género y se desarrollan en tres ambientes con rangos de nivel de ruido diferenciados según la actual legislación Colombiana sobre niveles máximos permisibles de ruido ambiental. Se resalta que el reconocimiento obtenido es independiente del hablante sin necesitar de los extensos entrenamientos previos que con otras herramientas se debe hacer.

Palabras clave: Microsoft SAPI, modelo de lenguaje, reconocimiento de voz, ruido ambiental, vocabulario cerrado.

ABSTRACT
This paper presents a computer application that recognizes Spanish voice command for a speaker independent closed vocabulary. The Spanish language model adopted is the one provided for Microsoft® SAPI (Speech Application Program Interface). This language model was limited to recognize only the grammar related with the functionalities that the user of the automated wheelchair studied by the Automatica research group of the Universidad Autónoma de Manizales can handle. The testing for measure the recognition system performance was implemented discriminately by gender and was developed in three environments with noise level ranges differentiated according the current Colombian legislation about maximum permissible ambient noise levels. It is highlighted that the recognition obtained is speaker independent without requiring the extensive previous training that with other tools should be done.

Keywords: closed vocabulary, environmental noise, language model, Microsoft SAPI, speech recognition.

1. INTRODUCCIÓN

La *Clasificación Internacional del Funcionamiento, de la Discapacidad y de la Salud* (CIF) define la discapacidad como un término genérico que engloba deficiencias, limitaciones de actividad y restricciones para la participación. La discapacidad denota los aspectos negativos de la interacción entre personas con un problema de salud y factores personales y ambientales (como actitudes negativas, transporte y edificios públicos inaccesibles, y falta de apoyo social) que impiden su participación plena y efectiva en la sociedad en pie de igualdad con los demás, ver [1, p. 4]. Como lo indica [1, p.xi] "en todo el mundo, las personas con discapacidad tienen peores resultados sanitarios y académicos, una menor participación económica y unas tasas de pobreza más altas que las personas sin discapacidad. En parte, ello es consecuencia de los obstáculos que entorpecen el acceso de las personas con discapacidad a servicios que muchos de nosotros consideramos obvios, en particular la salud, la educación, el empleo, el transporte, o la información".

En Colombia, el Ministerio de Salud y Protección Social, en [2], indica que la principal alteración permanente presentada en las personas incluidas en dicho registro, es la del "movimiento del cuerpo, manos, brazos y piernas", con un total de 394404 personas, equivalente al 33,5% del total del registro. La anterior alteración coincide con la principal dificultad permanente que las

personas incluidas en el RLCPD presentan en sus actividades diarias, la cual es: "caminar, correr, saltar", con un total de 591816 personas, equivalente al 50,2% del total de personas registradas; en esta misma estadística se indica que la dificultad para "llevar, mover, utilizar objetos con las manos" presenta un porcentaje del 18,7% equivalente a 220 244 personas. Según estadísticas anteriores, en Colombia, la discapacidad que más se presenta es la asociada con la discapacidad motora.

Para aquellas personas con discapacidad motora se han venido investigando e implementando diferentes métodos que les permitan tener control sobre su silla de ruedas. Actualmente es común encontrar en el comercio sillas de ruedas eléctricas controladas por *joystick*; por otra parte los estudios, especialmente, para pacientes que también carecen de la capacidad del habla, se han centrado en el uso de señales electromiográficas (EMG) tomadas en diversos músculos, [3], el uso de señales electro-oculográficas (EOG) con las que se puede detectar el movimiento de los ojos [4] y el uso de señales electroencefalográficas (EEG) que registran la actividad bioeléctrica cerebral, [5]. De igual manera se ha trabajado el control de movimiento de una silla de ruedas por medio de la detección de la dirección del rostro [6] y del movimiento de la lengua [7]; y para pacientes que sí se pueden comunicar oralmente, se trabaja también en el reconocimiento de comandos de voz, tema de interés en el presente artículo.

Particularmente, en el reconocimiento de comandos de voz, han sido varias las técnicas desarrolladas, como la DTW (Alineamiento temporal dinámico), cruce por cero, redes neuronales y HMM (Modelo Oculto de Markov), siendo esta última una de las más populares [8]. La implementación de las anteriores técnicas se puede encontrar en procesadores especializados de DSP (procesamiento digital de señales) para reconocimiento de voz, como lo es el "DSK TMS320C6711" de Texas Instruments, el "VR-Stamp" basado en el procesador RSC4128 y el kit de desarrollo para reconocimiento de voz "Voice Direct 364" ambos de Sensory Inc. En relación a Software, se encuentran diversos *toolkits* de código abierto para reconocimiento de voz, como [9], desarrollado por diferentes Instituciones de Japón; [10], desarrollado por la Universidad de Carnegie Mellon– EEUU; [11], desarrollado por la Universidad del Estado de Mississippi, y [12], especializado en HMM y desarrollado por Cambridge University. De igual manera la compañía Microsoft distribuye [13], una plataforma para el desarrollo del reconocimiento y síntesis de voz en su sistema operativo Windows.

Un proyecto de reconocimiento de voz similar al propuesto en este documento, [14], se desarrolló en la Universidad de Tottori en Japón. Los autores de este proyecto se apoyaron en la herramienta de software "Julian", la cual es otra versión de [9], para controlar una silla de ruedas por comandos de voz en idioma Japonés. Ellos adaptaron la silla de ruedas con un computador portátil en donde el proceso de reconocimiento fue llevado a cabo, obteniendo una tasa de reconocimiento exitoso del 98,3% para los comandos de movimiento. Con el mismo objetivo se encuentran desarrollos como [15], que realiza un control de voz basado en la nube para una silla de ruedas. Dicho control es implementado usando un *WebKit* con un API (Interfaz de Programación de Aplicaciones) de Voz en la nube, empleando librerías de Java Script que permiten el reconocimiento de voz y su conversión a texto y es soportado por el navegador de Google Chrome. Las pruebas se realizaron en idioma Ingles y Esloveno, utilizando 5 comandos compuestos, cada uno, por una sola palabra. Estas pruebas fueron realizadas en un entorno sin ruido de fondo, en el que 10 personas pronunciaron cada comando 15 veces, obteniendo un rango de precisión de reconocimiento que va desde 60% hasta 97%, aproximadamente. De manera similar se encuentra en [16] un desarrollo para reconocimiento en idioma inglés con la plataforma de software libre Pocketsphinx, que provee reconocimiento de voz continuo en tiempo real para dispositivos embebidos y requiere que los usuarios sean entrenados para pronunciar las palabras, como son esperadas por el Pocketsphinx. En sus pruebas, tres personas pronunciaron un conjunto de comandos que están formados tanto por palabras aisladas como por frases relacionadas con el movimiento de la silla, un total de 50 veces cada uno. El porcentaje de precisión del reconocimiento fue del 90% al 100% para la mayoría de los comandos. Otras opciones como el control de la rotación de una silla de ruedas por voz desde un télefono Android han sido exploradas en [17]. Además de lo anterior, se encuentran diversos artículos de desarrollos en control de sillas de ruedas por comandos de voz utilizando procesadores especializados de DSP existentes en el mercado para tal fin: [18; 19; 20; 21] con el que obtienen un reconocimiento de palabras aisladas, limitado por la capacidad de memoria del DSP y se hace necesario un entrenamiento previo por cada usuario para efectuar el reconocimiento.

En Colombia, se encuentran trabajos como [22], en el que se implementa el sistema de reconocimiento de voz sobre microcontroladores DSPIC's y se utiliza una red neuronal artificial como técnica de reconocimiento. El módulo reconoce 5 palabras del idioma español (adelante, atrás, izquierda, derecha y alto) y está entrenado para identificar palabras pronunciadas por un único hablante. Con este módulo se obtuvo un porcentaje de acierto en el reconocimiento del 90% para el promedio de las palabras pronunciadas. Otro trabajo a mencionar es [23], cuyo algoritmo fue implementado en Matlab, detectando palabras aisladas en un vocabulario pequeño, dependiente del locutor y en un ambiente controlado, con el que se obtuvo una eficiencia en el algoritmo del 96,08%. Similar al anterior se encuentra el proyecto [24], que también fue implementado en Matlab y permite la

identificación de comandos de voz con un diccionario reducido. Para este, se construyó una base de datos con los comandos: adelante, atrás, derecha, izquierda y pare. Los resultados obtenidos presentaron que la palabra con mayor cantidad de aciertos es *adelante*, con un porcentaje de 98%, y una dispersión del 2,4%. Por otra parte, la palabra que mayor dificultad presenta es *pare*, con un porcentaje de acierto de 87% y una dispersión de 9,3%.

El grupo de investigación en Automática de la Universidad Autónoma de Manizales UAM®, en la búsqueda por lograr disminuir las dificultades que en su desplazamiento deben afrontar las personas con discapacidad motora, viene liderando el proyecto integrador "Silla de ruedas automatizada". Como parte de las funciones de dicho proyecto se encuentra el reconocimiento de comandos de voz en español, con el que se puede favorecer a pacientes que tienen la capacidad del habla, pero que poseen dificultades importantes para desarrollar actividades que requieren la utilización de la destreza de los dedos de la mano para manipular por ejemplo botones o perillas, así como girar o torcer las manos o los brazos. Condiciones que no les permiten usar sus extremidades para hacer mover la silla de ruedas en la que se encuentran, ni accionar dispositivos de uso diario en su hogar o interactuar con un computador. Para esta clase de pacientes, el control de la silla de ruedas por comandos de voz es una opción cómoda, pues la voz sobresale como el medio de comunicación más natural y más usado para expresar lo que se desea.

Es así como en este artículo se presenta una aplicación computacional que reconoce comandos de voz en español para un vocabulario cerrado e independiente del hablante con una gramática enfocada a reconocer comandos relacionados con el movimiento de una silla de ruedas, con órdenes de domótica y con la toma de signos vitales. La interfaz gráfica es diseñada para guiar al usuario en los comandos a pronunciar, el desplazamiento entre las ventanas de la aplicación y el accionamiento de sus principales botones se puede controlar de igual manera por voz. Otras funcionalidades como el envío de correos electrónicos a destinatarios con su plantilla previamente almacenada por el usuario y la apertura de programas instalados en el computador también son implementadas por comandos de voz.

2. GENERALIDADES DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA Y PROBLEMAS EN LA DETECCIÓN

El propósito de un sistema de reconocimiento del habla es tomar como entrada la forma de onda acústica de la voz humana y producir como salida una cadena de palabras equivalente, [25]. Para lograr dicho resultado, la señal de voz ingresa a un módulo de procesamiento de señales en el que se extraen los vectores de características sobresalientes que son enviados posteriormente al decodificador; el decodificador utiliza tanto un modelo acústico como un modelo de lenguaje para generar finalmente la secuencia de palabras que tienen la máxima probabilidad de asemejarse a los vectores de características de entrada [8].

El modelo acústico se crea mediante la extracción de datos estadísticos de ficheros con voces recolectadas en el mismo idioma en el que se realizará el reconocimiento (corpus de habla-contiene los datos de una amplia población de oradores con su correspondiente transcripción). Esta información estadística es una representación del sonido que forma cada palabra. Mientras más información de voces se tenga, el modelo acústico será más exacto [26]. Este modelo incluye también información acerca de la acústica, la fonética, el micrófono y la variabilidad del medio ambiente, género y diferencias dialectales entre los hablantes, etc., como se observa en [27].

Por su parte, los modelos de lenguaje se refieren al conocimiento del sistema de lo que constituye una posible palabra; qué palabras tienen probabilidad de coocurrir y en qué secuencia, obteniendo así las probabilidades *a priori*, de las secuencias de palabras a reconocer. Para estimar este valor para secuencias de cualquier longitud se necesitaría una gran cantidad de datos, por lo que se debe acudir a aproximaciones. Una de las aproximaciones que están más extendidas son las basadas en N-Gramas. En estos tipos de modelos de lenguaje la probabilidad de aparición de una palabra únicamente depende de un número reducido de palabras que la preceden [28]. En un modelo 2-grama, por ejemplo (comúnmente llamado bigrama), la probabilidad de una palabra, dada la palabra anterior, se calcula como la frecuencia de secuencias de dos palabras, como por ejemplo "mover adelante" o "tomar presión". Estimadores como los N-gramas que asignan una probabilidad condicional a posibles próximas palabras se pueden utilizar para asignar una probabilidad conjunta para una frase entera [28; 25]. La terminología de asociar el valor *N* de un modelo N-grama con su orden, proviene de los modelos de Markov, en donde un modelo N-grama puede ser interpretado como un modelo de Markov de orden *N-1*, [6].

Estos sistemas de reconocimiento de voz se deben enfrentar a retos importantes relacionados con la gran cantidad de variables presentes en la señal de entrada. Una de ellas se asocia con las características del hablante (como lo son el estilo, tono y ritmo del habla, la fisiología, género, edad y acento), [27]. Se tiene que los patrones del habla de una persona pueden ser totalmente diferentes a los de otra, ya que estos dependen del tamaño físico de su tracto vocal, la longitud y anchura del cuello, que dependen en gran medida de la edad y el sexo y dan lugar a variaciones en la escala de frecuencias. También son importantes el estado de salud y su condición física (cansancio, gripa, etc.). Otras condiciones adversas importantes la constituyen el entorno y el canal de transmisión. El ruido de ambiente acústico suele

considerarse aditivo y es la más importante de las posibles condiciones adversas con que el reconocedor puede enfrentarse. También debe considerarse que el ruido puede estar presente desde el mismo dispositivo de entrada, como lo es el micrófono y ruidos de interferencia A/D (análogo a digital). El tipo y ubicación del micrófono pueden añadir ruido y distorsionar significativamente el espectro de la señal, [29; 23]. Otros factores que también afectan la señal son las interferencias y reverberaciones de la propia sala. Así como han de tenerse en cuenta las variaciones en el modo de articular del hablante debido a su reacción psicológica al entorno ruidoso, conocidos como efecto Lombard.

3. METODOLOGÍA

3.1 Plataforma en software seleccionada

La aplicación de reconocimiento de voz se desarrolló con el lenguaje de programación C#, utilizando Microsoft SAPI en un entorno de escritorio de Windows 7. La principal razón para seleccionar Microsoft SAPI se debe al modelo acústico que para el español posee Microsoft .NET Framework, el cual permite un reconocimiento independiente del hablante que se va entrenando conforme se va haciendo uso del mismo.

Con el espacio de nombres *System.Speech.Recognition* de Microsoft .NET Framework, se proporciona la funcionalidad para adquirir y monitorear la entrada de voz, crear gramáticas de reconocimiento del habla que produzcan tanto resultados de reconocimiento literales como semánticos, capturar información de eventos generados por el reconocedor de voz y configurar y administrar los motores de reconocimiento del habla [30]. La Gramática de Reconocimiento de habla se configuró con la clase *GrammarBuilder* que permite construir una gramática de un conjunto de frases y opciones.

Con respecto a las técnicas que utiliza Microsoft para el modelo acústico, se referencia en [31] el uso de un híbrido entre un pre-entrenamiento de redes neuronales profundas (DNN) y un modelo oculto de Markov (HMM) dependiente del contexto (CD) para el reconocimiento del habla en vocabulario largo, técnica reconocida con la abreviación CD-DNN-HMMs. Esta arquitectura híbrida entrena las redes neuronales profundas para producir una distribución sobre senones (estados de trifonemas atados) como sus salidas. Según [31], el entrenamiento de redes neuronales para predecir una distribución sobre senones brinda mayor cantidad de bits de información que estarán presentes en las etiquetas de la red neuronal entrenada.

3.2 Modelo de lenguaje para la aplicación desarrollada bajo un sistema de reconocimiento con vocabulario cerrado

Se adaptó el modelo de lenguaje de propósito general para el español del SAPI de Microsoft a las necesidades específicas de la aplicación, en donde solo se requiere reconocer ciertas expresiones de interés para la misma (vocabulario cerrado). Por lo tanto, se definió una gramática que limita el reconocedor para escuchar solo el habla que le interesa a la aplicación. Con un vocabulario cerrado se obtienen beneficios como los mencionados en [32]:

- Se aumenta la precisión y rendimiento del reconocedor comparado con tareas de dictado (vocabulario abierto).
- Se garantiza que todos los resultados del reconocimiento tengan significado para la aplicación, y permite al motor de reconocimiento especificar los valores semánticos inherentes en el texto reconocido.
- Reduce la sobrecarga de procesamiento que la aplicación requiere.
- Permite un procesamiento independiente del locutor, lo que elimina la necesidad de entrenar el reconocedor para configurar perfiles por cada hablante.

En el desarrollo de la aplicación, las clases del espacio de nombres *System.Speech.Recognition* utilizadas en la construcción de la gramática para los comandos seleccionados son:

- **Choices**: Representa una lista de alternativas posibles que el usuario pronunciará dentro de las restricciones de una gramática de reconocimiento de voz.
- **GrammarBuilder**: Proporciona un mecanismo para construir las restricciones de una gramática de reconocimiento de voz, permitiendo armar una gramática a partir de un conjunto de frases y opciones (*Choices*). De esta manera se puede definir la forma en que las palabras pueden ser combinadas para ser entendidas por el reconocedor.
- **Grammar**: Proporciona soporte en tiempo de ejecución para la obtención y gestión de la información de una gramática de reconocimiento de voz.

Las gramáticas prefijadas para los diferentes comandos que contiene la aplicación corresponden a los relacionados con el movimiento de la silla, a órdenes de domótica, a la toma de signos vitales, al desplazamiento por las pestañas de la aplicación y a la activación de sus principales botones. De la figura 1 a la figura 4 se visualizan algunos de los comandos a pronunciar según gramáticas prefijadas. Se tiene también otra gramática de libre configuración por el usuario, que se actualiza en tiempo de ejecución y se asocia a la correspondencia de un comando con un destinatario de correo electrónico y al control de la apertura de aplicaciones predefinidas instaladas en el computador del usuario.

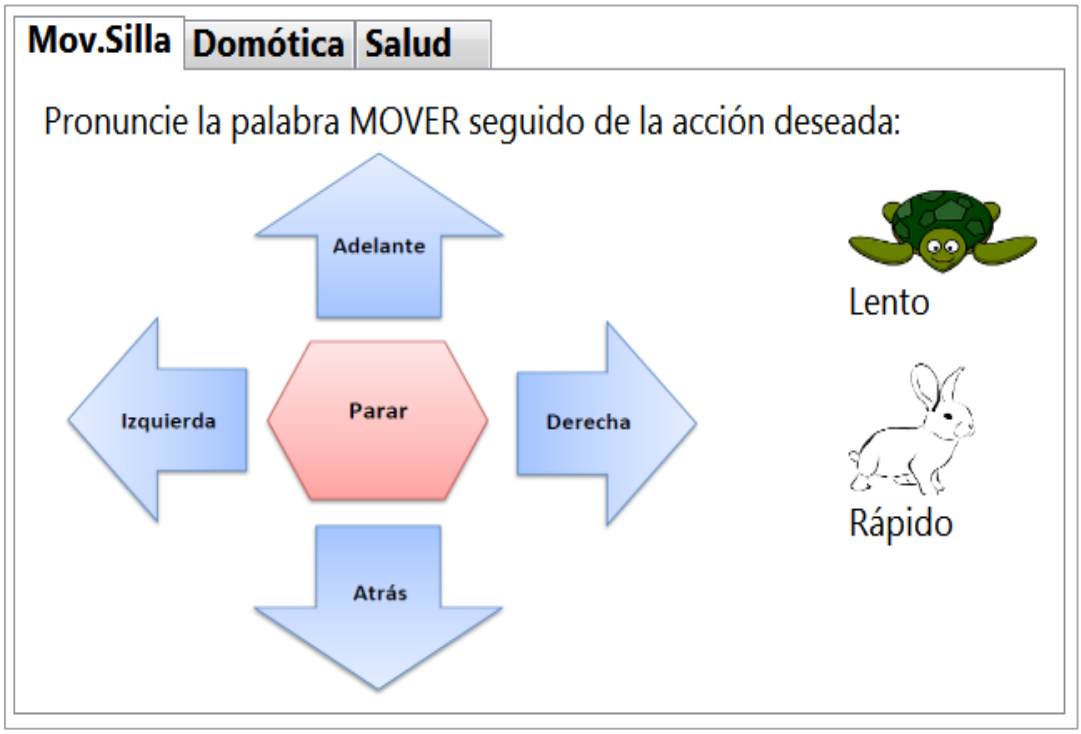

Figura 1. Comandos referentes al movimiento de la silla. Fuente: Elaboración propia.

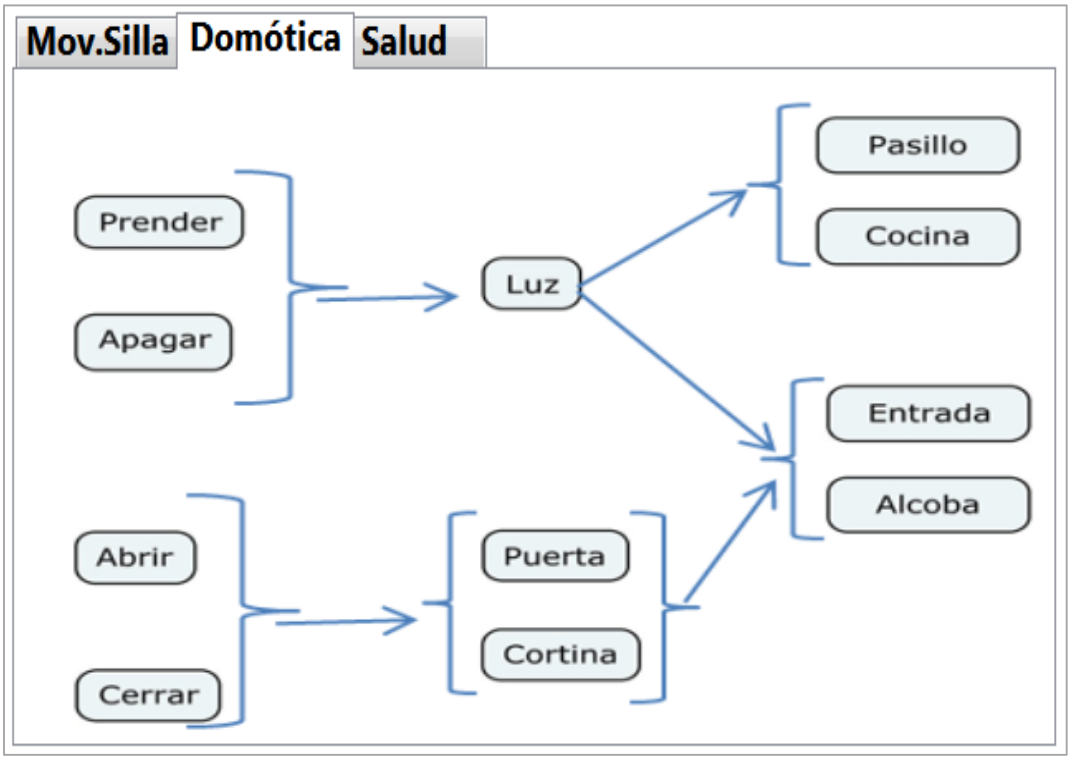

Figura 2. Secuencia de palabras para formar frases relacionadas con órdenes de domótica. Fuente:Elaboración propia.

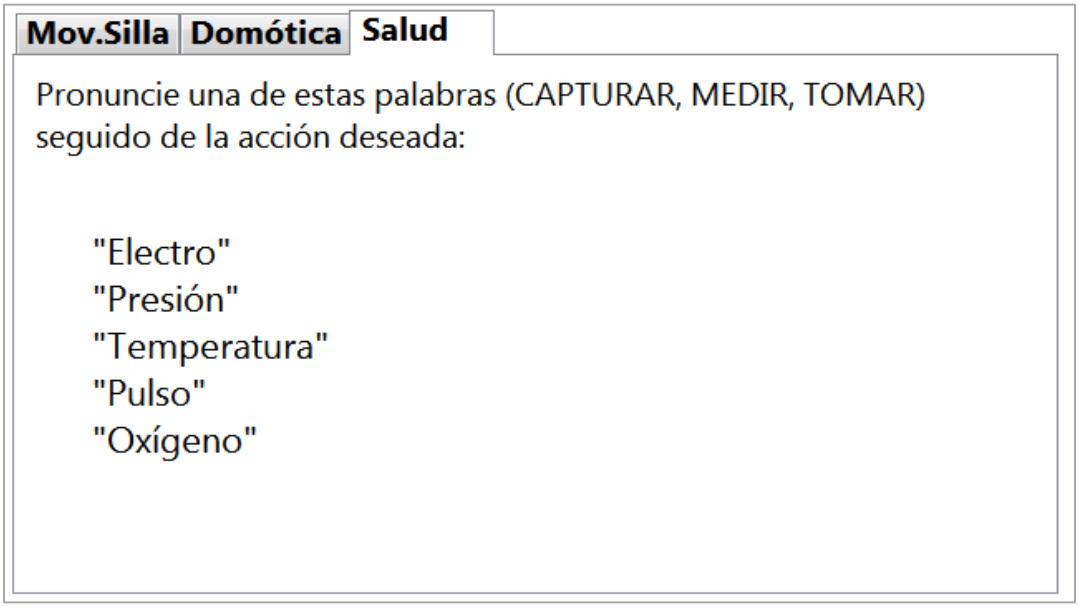

Figura 3. Comandos referentes a la toma de signos vitales. Fuente: Elaboración propia..

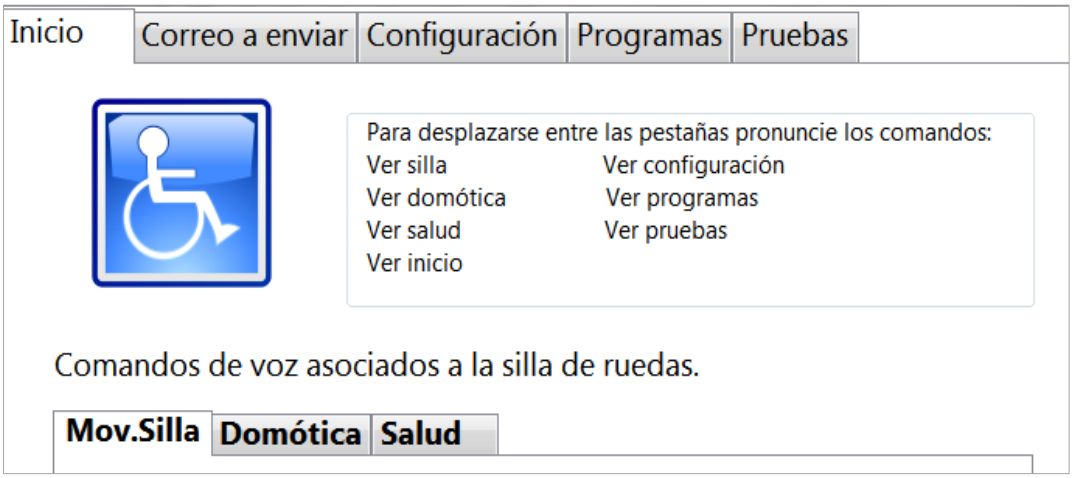

Figura 4. Comandos referentes al desplazamiento por las pestañas de la aplicación. Fuente: Elaboración propia

Las figuras 5 y 6 muestran las gramáticas configurables por el usuario. Para cargar en la aplicación cada una de las gramáticas definidas, se utiliza el método LoadGrammar**,** que realiza la carga sincrónicamente. Este método pertenece a la clase *SpeechRecognition Engine*, la cual, a su vez, pertenece al espacio de nombre *System.Speech.Recognition*.

La aplicación se desarrolló bajo una interfaz compuesta principalmente por pestañas a las que se puede acceder por comandos de voz. El aspecto general de la interfaz se visualiza en la figura 7. En la misma, la sección fija (lado derecho) contiene información de interés referente al estado de la entrada de audio y al resultado del reconocimiento

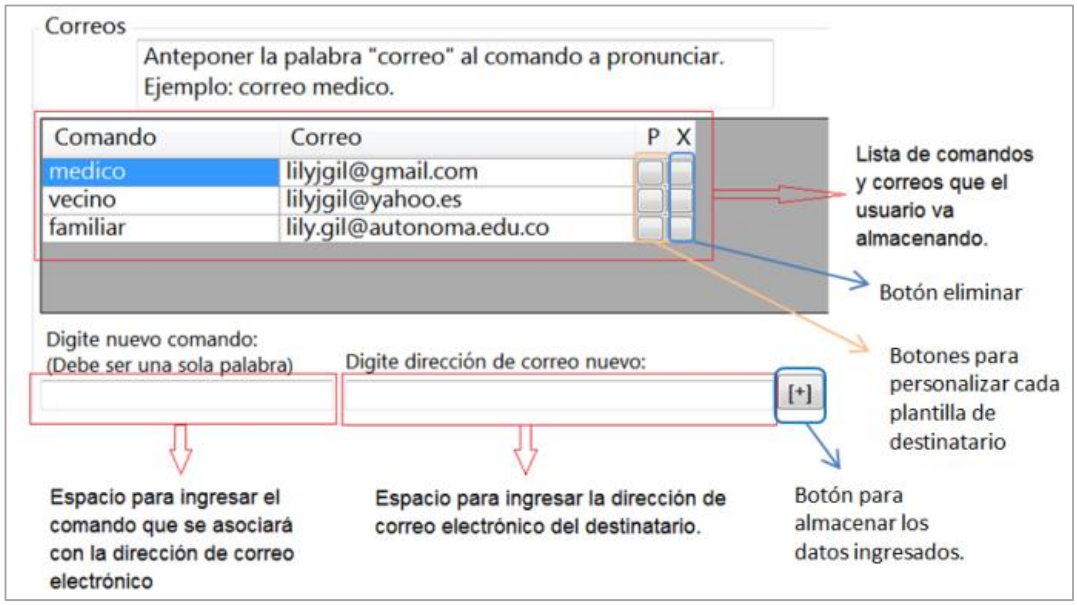

Figura 5. Entorno en el que el usuario configura los comandos asociados a cuentas de correo electrónico de destinatarios deseados. Fuente. Elaboración propia.

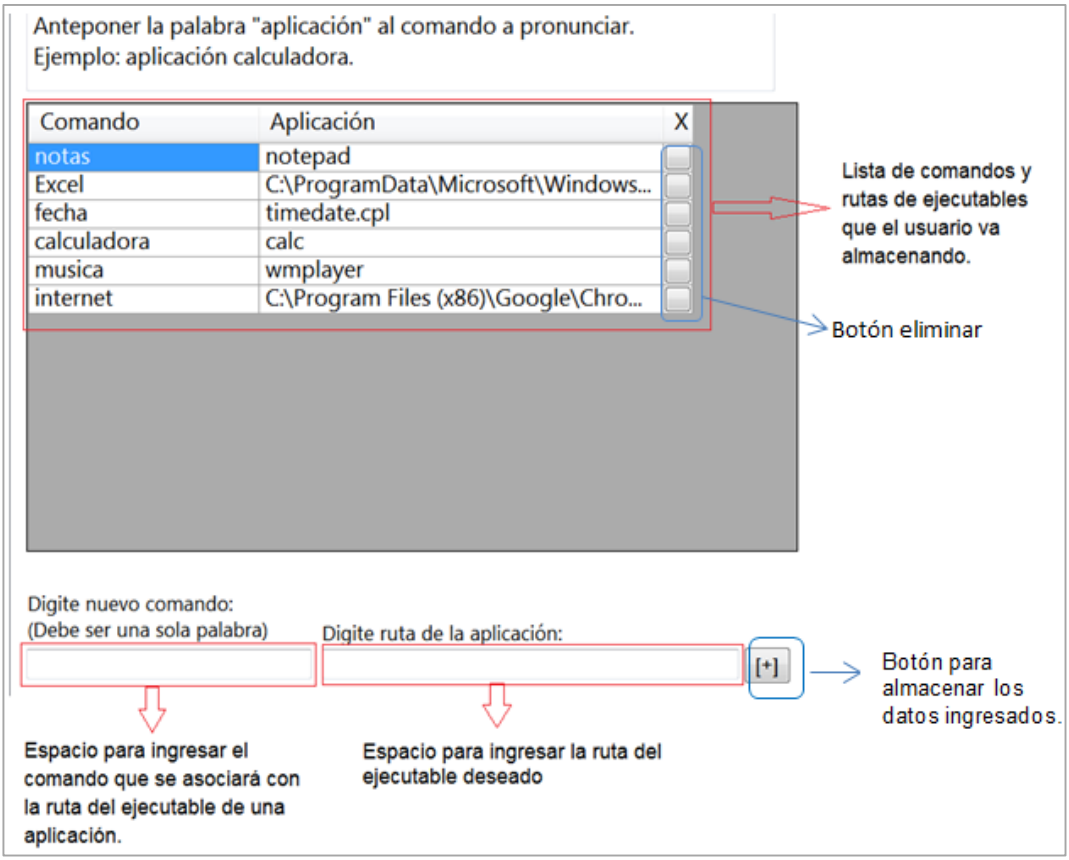

Figura 6. Entorno en el que el usuario configura el comando asociado a una ruta del ejecutable de una aplicación deseada. Fuente. Elaboración propia.

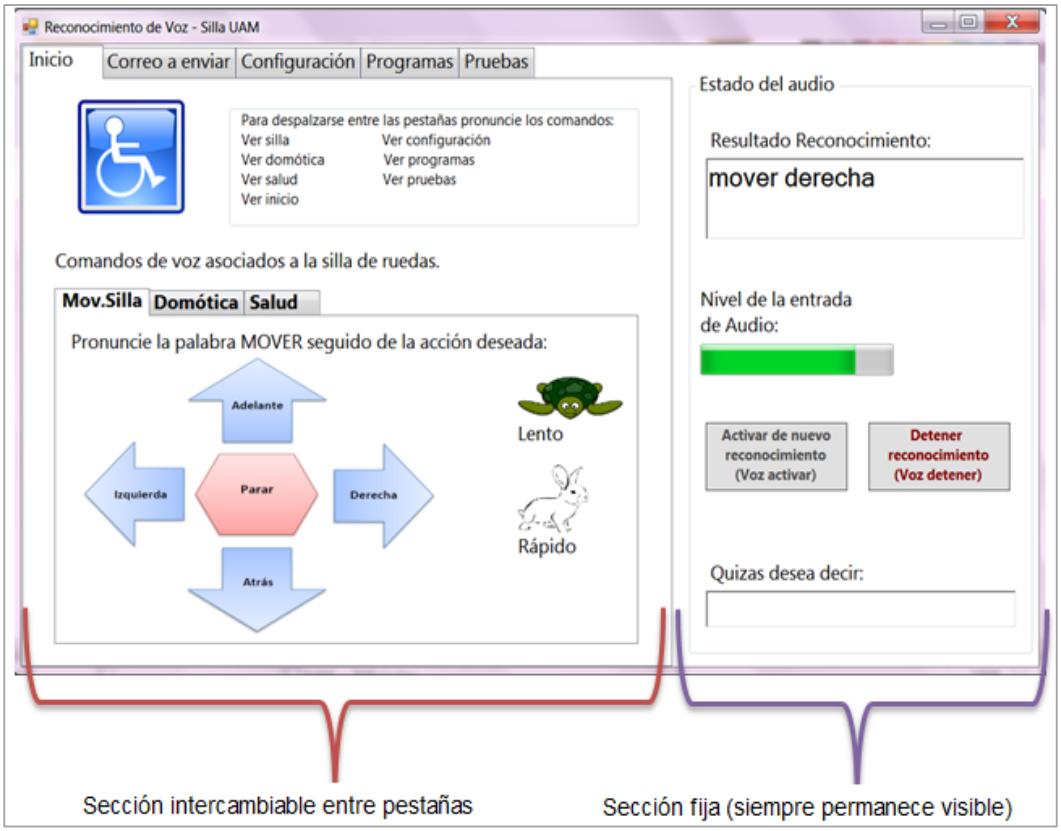

Figura 7. Aspecto general de la interfaz. Fuente. Elaboración propia.

4. PRUEBAS Y RESULTADOS

4.1 Pruebas para determinar el nivel de desempeño obtenido en la aplicación

Para determinar el nivel de desempeño obtenido en la aplicación para el reconocimiento de los comando de voz en español en un vocabulario cerrado e independiente del hablante, se realizaron pruebas con 10 hombres y 10 mujeres de nacionalidad colombiana, quienes debían colocarse un micrófono diadema a una distancia de aproximadamente 3 cm de la boca, al momento de pronunciar los comandos.

Por cada participante se involucraron tres rangos de nivel de ruido ambiental de acuerdo con la vigente resolución 0627 del 7 de Abril de 2006 del entonces Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MAVDT), por la cual se establece la Norma Nacional De Emisión De Ruido y Ruido Ambiental, en cuyo capítulo III denominado "Del ruido ambiental", se indican los estándares máximos permisibles de niveles de ruido ambiental, expresados en decibeles dB(A). El decibelio con ponderación A (dB(A)) es una unidad de nivel sonoro medido con un filtro previo que quita parte de las frecuencias muy bajas y muy altas, adaptándose a la percepción del oído humano, siendo la unidad más utilizada para la medición de los niveles de ruido ambiental [33].

Para asegurar que las pruebas con todos los usuarios se realizaran dentro del mismo rango de dB(A) se utilizó como instrumento de medida un sonómetro marca UNI-T, referencia UT352, el cual tiene un rango de medición entre 30dB y 130dB, con una exactitud de ±1.5dB. El sonómetro se configuró para medir con el filtro de ponderación frecuencial A y el filtro de ponderación temporal F (Rápido), que tiene un tiempo de respuesta de 125 ms. Esta medición se efectuó justo junto al micrófono que el usuario, por medio de una diadema, ya tenía ubicado cerca de su boca, tomando así el valor de dB(A) que se estaba percibiendo alrededor del micrófono.

La primea prueba se realizó en un espacio cerrado, silencioso y alejado del tráfico vehicular, con mediciones en el sonómetro que se encuentran dentro del rango de un "sector A: Tranquilidad y Silencio", según los estándares máximos permisibles de niveles de ruido ambiental. Para las pruebas dos y tres se adicionó ruido al lugar con la ayuda de la aplicación en línea *myNoise™.net*, la cual cubre todo el rango de frecuencia audible desde los 20 Hz hasta los 20 KHz, seleccionándose en el mismo el generador de ruido de fondo llamado *Coffee-Shop* que

simula el ruido que se genera en una cafetería concurrida, donde hay sonido de cubiertos, de objetos retumbando, de personas charlando, murmurando y tosiendo, entre otros ruidos. Además, se conectó un parlante externo que amplificó los niveles de decibeles requeridos. Las pruebas dos y tres se realizaron dentro del rango para un "Sector C: Ruido Intermedio Restringido", según los estándares máximos permisibles de niveles de ruido ambiental y, específicamente, la prueba tres incluyó mediciones superiores a los 80dB(A), que es el valor máximo permisible de nivel de ruido ambiental que se encuentra en la legislación y pertenece al caso de zonas al aire libre, como parques mecánicos, áreas destinadas a espectáculos públicos, vías troncales, autopistas, vías arterias, vías principales, entre otras. La tabla I muestra los rangos definidos para los tres tipos de pruebas realizadas.

TABLA I
RANGOS DE RUIDO ESTABLECIDOS PARA LAS TRES PRUEBAS

	Rango dB(A)	Ruido
Prueba # 1	35 dB(A) hasta 55 dB(A)	Lugar en silencio
Prueba # 2	60 dB(A) hasta 72 dB(A)	Adicionando ruido tipo Coffee-Shop.
Prueba # 3	73 dB(A) hasta 85 dB(A)	

Fuente: Elaboración propia.

Cada persona pronunció por cada una de las tres pruebas 35 comandos, repitiendo cada uno de los mismos cuatro veces. De tal manera que se pronuncian en total 140 comandos en cada prueba por persona. Antes de empezar se le aclaró a cada participante que debía pronunciar los comandos con la misma intensidad para los tres rangos de nivel de ruido, sin subir la voz en las pruebas dos y tres, en las que el ruido era mayor. Esto, para evitar efecto Lombard. El análisis de los resultados obtenidos en las pruebas se realizó por medio de una matriz de confusión para cada variable a analizar: sexo y nivel de ruido. Dicha matriz es una herramienta estadística de visualización que permite evaluar la eficiencia del sistema de reconocimiento.

Los resultados en un entorno en silencio (prueba #1) con rango de nivel de ruido de 35 dB(A) hasta 55 dB(A) para ambos géneros tuvieron el mismo comportamiento, obteniéndose un reconocimiento exitoso del 100% de los comandos pronunciados, sin presentarse casos de omisión o de sustitución entre los mismos. La tabla II resume los parámetros de eficiencia calculados sobre las matrices de confusión resultantes.

TABLA II
PARÁMETROS DE EFICIENCIA CALCULADOS SOBRE LAS MATRICES DE CONFUSIÓN EN LA PRUEBA # 1 EN HOMBRES Y EN MUJERES

Parámetro	Prueba #1 35 dB(A) hasta 55 dB(A)	
	Mujeres	**Hombres**
Exactitud	100%	100%
Sensibilidad	100% en todos los comandos	100% en todos los comandos
Especificidad	100% en todos los comandos	100% en todos los comandos
Precisión	100% en todos los comandos	100% en todos los comandos
Medida F1	100% en todos los comandos	100% en todos los comandos

Fuente: Elaboración propia.

Los resultados en la prueba #2 para ambos géneros, donde el nivel de ruido se controló para que permaneciera entre los 60 dB(A) hasta los 72 dB(A), tuvieron un comportamiento casi igual. Solo se presentó error con uno de los comandos pronunciados. Para el caso de las mujeres, el comando "ver configuración" fue reconocido en una oportunidad como "tomar presión", por lo que se dio un error de sustitución. Por su parte, para el caso de los hombres, el comando "ver inicio" no fue reconocido en una oportunidad y no se relacionó con algún otro comando, obteniéndose así un error de omisión. Los parámetros de eficiencia calculados sobre las matrices de confusión y presentados en porcentajes, se resumen en la tabla III.

Para la prueba #3 donde el nivel de ruido se controló para que permaneciera entre los 73 dB(A) hasta los 85 dB(A), se obtuvo como resultado, para el caso de las mujeres, que 28 de los 35 comandos pronunciados tuvo un reconocimiento exitoso del 100%. Los siete comandos restantes solo evidenciaron errores de omisión, al no ser identificados ni reconocidos como otro comando, sin presentarse por lo tanto falsos positivos. Por su parte, en la prueba con hombres, 21 de los 35 comandos pronunciados obtuvo un reconocimiento exitoso del 100%. Los 14 comandos restantes mostraron errores, ya sea de omisión o de sustitución. Los parámetros de eficiencia calculados sobre las matrices de confusión y presentándose en porcentaje se resumen en la tabla IV.

TABLA III
PARÁMETROS DE EFICIENCIA CALCULADOS SOBRE LAS MATRICES DE CONFUSIÓN EN LA PRUEBA # 2 EN HOMBRES Y EN MUJERES

Parámetro	Prueba #2 60 dB(A) hasta 72 dB(A)	
	Mujeres	Hombres
Exactitud	99,93%	99,93%
Sensibilidad	97,5% en solo uno de los comandos. El resto 100%	97,5% en solo uno de los comandos. El resto 100%
Especificidad	99,93% en solo uno de los comandos. El resto 100%	100% en todos los comandos
Precisión	97,56% en solo uno de los comandos. El resto 100%	100% en todos los comandos
Medida F1	Superior al 98,7% en dos de los comandos. El resto 100%	98,73% en solo uno de los comandos. El resto 100%

Fuente. Elaboración propia.

TABLA IV
PARÁMETROS DE EFICIENCIA CALCULADOS SOBRE LAS MATRICES DE CONFUSIÓN EN LA PRUEBA # 3 EN HOMBRES Y EN MUJERES

Parámetro	Prueba #3 73 dB(A) hasta 85 dB(A)	
	Mujeres	Hombres
Exactitud	99,36%	98,29%
Sensibilidad	Superior al 94,99% en siete de los comandos. El resto 100%	Superior al 92,4% en catorce de los comandos. El resto 100%
Especificidad	100% en todos los comandos	Superior al 99,8% en tres de los comandos. El resto 100%
Precisión	100% en todos los comandos	Superior al 95,2% en tres de los comandos. El resto 100%
Medida F1	Superior al 97,4% en siete de los comandos. El resto 100%	Superior al 96% en diez y siete de los comandos. El resto 100%

Fuente. Elaboración propia.

Según resultados, de la tabla II a la tabla IV, el sistema de reconocimiento de voz en español para un vocabulario cerrado e independiente del hablante, no presenta diferencias significativas en su desempeño al responder ante hombres y mujeres. Solo en la prueba #3, la respuesta presenta en el caso de las mujeres valores levemente superiores en todos los parámetros de eficiencia que los obtenidos por los hombres. Así mismo, el sistema de reconocimiento responde en generar muy bien en los tres ambientes de prueba, dándose una leve desmejora a medida que el ruido en el ambiente aumenta.

4.2 Observación respecto a la influencia de subir la voz al momento de pronunciar los comandos

La tendencia involuntaria a incrementar el esfuerzo vocal cuando se habla en un lugar ruidoso con el fin de mejorar la audibilidad de la voz se conoce como efecto Lombard e interfiere enormemente en la respuesta del reconocedor, ya que los cambios al subir la voz afectan no solo a la sonoridad, sino también a factores como el tono, el rango y la duración del sonido de las sílabas. Cuando un locutor habla en presencia de ruido, estudios como [29] han encontrado que el primer formante de una vocal tiende a crecer mientras que el segundo decrece y que la caída espectral decrece en las frecuencias bajas y aumenta en las altas para la mayoría de las vocales.

Para comprobar el efecto de subir la voz al momento de pronunciar los comandos en un ambiente ruidoso se realizó una prueba con los mismos rangos de nivel de ruido de la prueba #3 (entre 73 dB(A) hasta 85 dB(A) con tres de los participantes que de igual manera debían repetir cada comando cuatro veces, y se analizaron los resultados por medio de una matriz de confusión. En esta prueba, el valor de exactitud de la matriz de confusión bajó a un 55,77 %, cuando en las pruebas anteriores todos los resultados habían sido mayores al 98%. Solo el 7,7% de los comandos obtuvo un valor de sensibilidad del 100%. El 35,9% de los mismos obtuvo un valor inferior al 50% y el 56,4% obtuvo un valor entre el 50% y el 91,7%. El valor de especificidad y de precisión fue del 100% para el 79,5% de los comandos, siendo el comando "tomar electro" el que obtuvo el valor más bajo tanto de especificidad como de precisión, con un valor de 84,2%, y 19% respectivamente.

4.3 Prueba para establecer el valor de confidencia adecuado para los comandos de primer nivel

El valor de confidencia es un parámetro que da una restricción de nivel de confianza al reconocedor. Si el valor es muy bajo, puede detectar erróneamente palabras pronunciadas que no están en el vocabulario como válidas y si es muy alto puede bloquear una mayor cantidad de frases que sí son correctas y tomarlas como no válidas. El rango en el que se puede fijar el nivel de confidencia va entre 0 (mínimo) y 1(máximo). Con el fin de establecer un valor de confidencia adecuado para los comandos fijos que componen el vocabulario cerrado de la aplicación, se realizó una prueba que incluye las 13 clases de primer nivel. De cada una de estas 13 clases se desprenden las diferentes frases que componen los comandos de la aplicación.

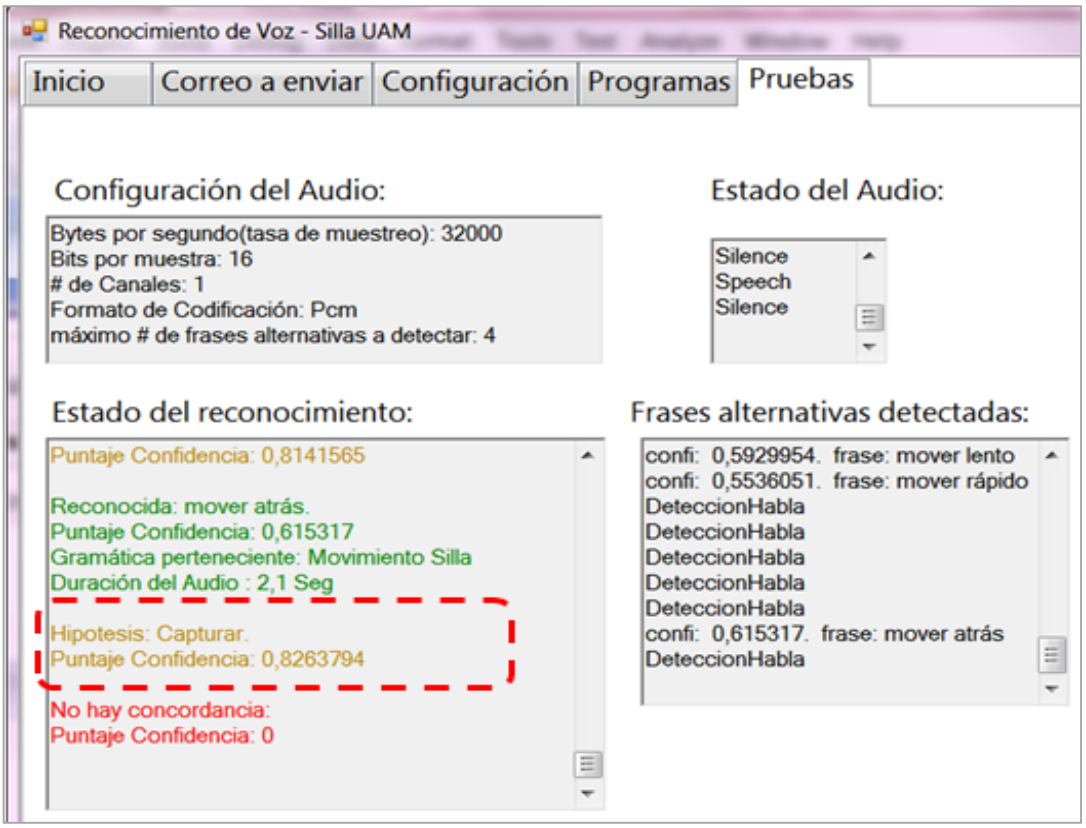

Figura 8. Monitoreo del comportamiento del reconocedor de voz. Fuente. Elaboración propia.

La tabla V muestra la palabra que corresponde a cada clase. De forma aleatoria, cada una de ellas se pronunció en un total de 20 veces, anotando el valor de confidencia con que eran reconocidas según el monitoreo del comportamiento del reconocedor de voz, visualizado en la pestaña "Pruebas" de la aplicación, como se muestra en la figura 8.

La figura 9 muestra la media del valor de confidencia para cada una de las 13 clases de la tabla V, mientras que la figura 10 muestra los resultados del valor de confidencia en una gráfica tipo boxplot.

De acuerdo con los datos registrados en la figura 9, el 76,9% de las clases obtuvo un valor de confidencia con una media por encima de 0,7, la clase 12 ("ver") presentó el valor de media más bajo con 0,604. Por otra parte, la figura 10 muestra que de las 20 veces que se pronunció cada palabra, la clase que presentó menor dispersión en la distribución del valor de confidencia aceptado fue la clase 3 ("aplicación"), con una diferencia inferior a 0,058 entre el menor y el mayor de los valores. La clase 12 ("ver") presentó el valor de confidencia más bajo de todos los casos, con 0,3265 y las clases 10, 6 y 5 ("prender", "correo" y "cerrar", respectivamente) presentaron el valor de confidencia más alto, siendo 0,8565 el mismo valor para las tres.

TABLA V
CLASES DE PRIMER NIVEL

Clase	Comando de primer nivel	Clase	Comando de primer nivel
Clase 1	'abrir'	Clase 8	'medir'
Clase 2	'apagar'	Clase 9	'mover'
Clase 3	'aplicación'	Clase 10	'prender'
Clase 4	'capturar'	Clase 11	'tomar'
Clase 5	'cerrar'	Clase 12	'ver'
Clase 6	'correo'	Clase 13	'voz'
Clase 7	'enviar'		

Fuente. Elaboración propia.

Según resultados anteriores, se puede fijar un valor de confidencia de 0,6 al momento de configurar la restricción de aceptación del reconocedor, valor superado en la media de todas las clases de la tabla V. Con esto que se pretende que los comandos pronunciados válidos sean aceptados como tal, pero que a la vez exista un nivel de rechazo para los casos en los que hay comandos supuestamente reconocidos pero que por su bajo nivel de

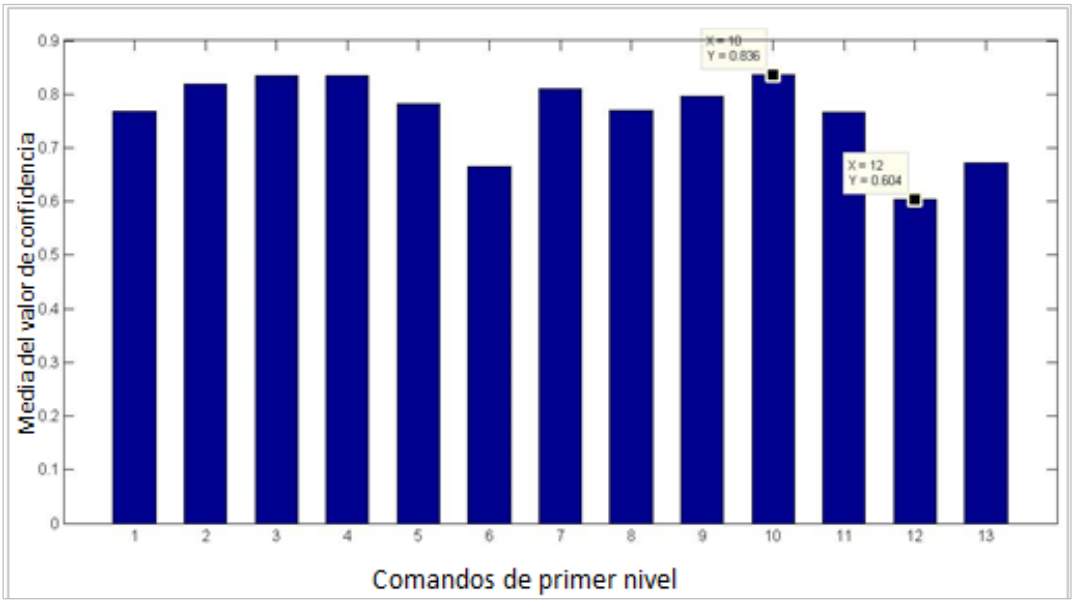

Figura 9. Media del valor de confidencia para las 13 clases de primer nivel. Fuente: Elaboración propia.

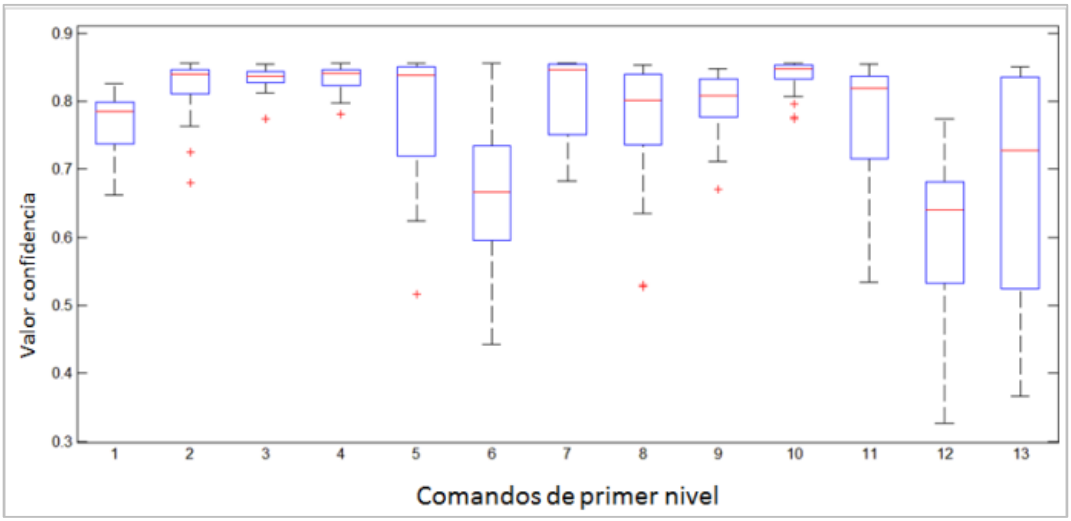

Figura 10. Distribución del valor de confidencia para las 13 clases de primer nivel. Fuente: Elaboración propia

confianza generen gran incertidumbre con relación a la precisión y asertividad del proceso de reconocimiento.

Es también importante destacar de los resultados que las palabras 'voz', 'ver' y 'correo' pueden presentar mayores errores al momento de ser reconocidas, por su gran dispersión (ver figura 10), porque el 25% de los datos en cada una de ellas están por debajo de un valor de confidencia de 0,6 y porque son las palabras que presentaron los casos con valores de confidencia más bajos.

5. CONCLUSIONES

Se implementó una interfaz gráfica que permite observar la retroalimentación del comportamiento del sistema, informando de manera visual, al momento de pronunciar las frases, si el comando es reconocido y en caso de serlo muestra cual es el comando. Así mismo, la activación o suspensión del sistema de reconocimiento se puede controlar por comando de voz, aspecto clave para evitar que la aplicación reconozca comandos cuando no se les están dictando.

El SAPI de Microsoft tiene ya muy desarrollado un modelo de lenguaje para el idioma español, modelo que se adaptó a las necesidades específicas de la aplicación, en la cual se limitó el vocabulario a comandos compuestos por dos o más palabras en un orden específico, relacionado con las funcionalidades de la silla de ruedas Automatizada de la UAM®. La aplicación demostró ser independiente del hablante y no requerir de entrenamientos previos, puesto que cada persona para realizar las pruebas solo tuvo que empezar a pronunciar los comandos definidos e inmediatamente el sistema los empezó a reconocer exitosamente.

Se validó la respuesta del sistema de reconocimiento de comandos de voz en español, visualizando los resultados en matrices de confusión sobre las que se calcularon parámetros de eficiencia correspondientes a la exactitud global, a la sensibilidad, la especificidad, la precisión y la medida F1 de los diferentes comandos pronunciados en las pruebas. Según los resultados encontrados, no hay diferencias significativas en la respuesta del sistema según género del interlocutor. Por su parte, al realizar el análisis en los tres rangos de nivel de ruido se encontró que a medida que el ruido aumenta, la respuesta del sistema de reconocimiento va disminuyendo en muy poca proporción. Los errores sobre los comandos se presentaron en mayor medida por omisión, que por sustitución entre los mismos.

6. RECOMENDACIONES

Se debe evitar subir la voz al momento de pronunciar los comandos en entornos con nivel de ruido apreciable (efecto Lombard), ya que esto interfiere enormemente en la respuesta del reconocedor, dando como resultado una disminución importante en todos los parámetros de eficiencia del mismo y aumentando los casos de errores por omisión o por sustitución entre los comandos.

Para conformar los diferentes comandos se recomienda utilizar palabras que tengan más de una sílaba y que no presenten problemas comunes en la pronunciación como es el caso de la doble *r*, ya que este tipo de palabras obtuvo una mayor probabilidad de reconocimiento erróneo, así como los casos con valores de confidencia más bajos, según prueba para establecer el valor de confidencia adecuado para los comandos de primer nivel.

7. AGRADECIMIENTOS

Se agradece el apoyo recibido por parte de los docentes del departamento de Electrónica y Automatización y del departamento de Ciencias Computacionales de la Universidad Autónoma de Manizales.

8. REFERENCIAS

[1] Organización Mundial de la Salud y Banco Mundial. (2011) *Informe mundial sobre la discapacidad*. [En línea]. Disponible en: https://goo.gl/0KtNAI

[2] Ministerio de Salud y Protección. (2015). *Registro para la localización y caracterización de personas con discapacidad (RLCPD)*".

[3] C.S.L. Tsui et al, "EMG-based hands-free wheelchair control with EOG attention shift detection," en *IEEE Int'l Conf. Robotics and Biomimetics (ROBIO 2007)*, dic. 15-18, 2007, pp. 1266-1271. DOI: 10.1109/ROBIO.2007.4522346

[4] S. Yathunanthan et al, "Controlling a Wheelchair by Use of EOG Signal," en *4th Int'l Conf. Information and Automation for Sustainability (ICIAFS 2008)*, dic. 12-14, 2008, pp. 283-288. DOI: 10.1109 /ICIAFS.2008.4783987

[5] I. Iturrate, J. Antelis y J. Minguez, "Synchronous EEG brain-actuated wheelchair with automated navigation," en *IEEE Int'l Conf. Robotics and Automation (ICRA '09)*, may. 12-, 2009, pp. 2318-2325. DOI: 10.1109/ROBOT.2009.5152580

[6] Z. Hu et al., "A novel intelligent wheelchair control approach based on head gesture recognition," en *Int. Conf. Computer Application and System Modeling (ICCASM)*, oct. 22-24, 2010, pp. V6-159-V6-163. DOI: 10.1109/ICCASM.2010.5619 307

[7] M.E. Lund et al, "Inductive tongue control of powered wheelchairs," en *Annual International Conference of the IEEE. Engineering in Medicine and Biology Society (EMBC)*, ago. 31, 2010-sep. 4, 2010, pp. 3361-3364. DOI: 10.1109/IEMBS.2010 .5627923

[8] X. Huang y L. Deng, "An Overview of Modern Speech Recognition," en *Handbook of Natural Language Processing*, 2a ed.: Chapman & Hall/CRC, 2010, ch. 15 (ISBN: 1420085921), pp. 339-366.

[9] Julius (2014) *Open-Source Large Vocabulary CSR Engine Julius*. [En línea]. Disponible en: http://julius.sourceforge.jp/en_index.php?q=index-en.html

[10] CMU (2016) *CMU Sphinx-Open Source Toolkit*. [En línea]. Disponible en: http://cmusphinx .sourceforge.net/

[11] The Institute for Signal and Information Processing. (2016) *ISIP toolkit. About our software*. [En línea]. Disponible en: http://www.isip.piconepress.com /projects/speech/software/

[12] (2016) *HTK Speech Recognition Toolkit*. [En línea]. Disponible en: http://htk.eng.cam.ac.uk/

[13] Microsoft (2016) *Microsoft Developer Network. Speech API.* [En línea]. Disponible en: https://goo.gl/XIc7po

[14] M. Nishimori, T. Saitoh y R Konishi, "Voice controlled intelligent wheelchair," en *SICE, 2007 Annual Conference*, Takamatsu, 2007, pp. 336-340. DOI: 10.1109/SICE.2007.4421003.

[15] A. Škraba et al, "Speech-controlled cloud-based wheelchair platform for disabled persons," *Microprocessors and Microsystems*, vol. 39, num. 8, nov.2015, pp. 819-828. DOI: 10.1016/j.micpro .2015.10.004

[16] J.A. Ansari, A. Sathyamurthy y R. Balasubramanyam, "An Open Voice Command Interface Kit," en *IEEE Transactions on Human-Machine Systems*, vol. 46, num. 3, jun. 2016, pp. 467-473, DOI: 10.1109/THMS.2015.2476458.

[17] S.U. Khadilkar y N. Wagdarikar, "Android phone controlled voice, gesture and touch screen operated smart wheelchair," en *International Conference on Pervasive Computing (ICPC)*, Pune, 2015, pp. 1-4. DOI:10.1109/PERVASIVE.2015.7087119.

[18] M. Fezari y A. Khati, "New speech processor and ultrasonic sensors based embedded system to improve the control of a motorised wheelchair," en *3rd International Design and Test Workshop (IDT)*, dic. 20-22, 2008, pp. 345-349. DOI: 10.1109/IDT.2008.4802527

[19] M.T. Qadri y S.A. Ahmed, "Voice Controlled Wheelchair Using DSK TMS320C6711," en *Int. Conf. on Signal Acquisition and Processing. (ICSAP)*, abr. 3-5, 2009, pp. 217-220. DOI: 10.1109/ICSAP.2009.48

[20] M. Fezari, M. Bousbia-Salah y M. Bedda, "Voice and Sensor for More Security on an Electric Wheelchair," en *2nd Int. Conf. on Info. and Comm. Tech. (ICTTA)*, 2006, pp. 854-858. DOI: 10.1109/ICTTA.2006.1684485

[21] C. Aruna et al, "Voice recognition and touch screen control based wheel chair for paraplegic persons," en *International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, mar. 6-8, 2014, pp. 1-5. DOI: 10.1109/ICGCCEE.2014.6922215

[22] J.C. Martínez y J.L. Ramírez, "Diseño y construcción de un módulo automático controlado por voz adaptable a una silla de ruedas convencional," *Segundo Congreso Internacional de Ingeniería Mecatrónica*, vol. 1, num. 1, pp. 1234-1234, Colombia, 2009.

[23] O.I. Higuera, "Diseño e implementación de un prototipo de reconocimiento de voz basado en modelos ocultos de markov para comandar el movimiento de una silla de ruedas en un ambiente controlado," en *XII Simposio de Tratamiento de Señales, Imágenes y Visión artificial*, Colombia, 2007.

[24] W. Acosta, M. Sarria y L. Duque, "Implementación de una metodología para la detección de comandos de voz utilizando HMM," *Revista de Investigaciones Universidad del Quindío*, vol. 23, num. 1, pp. 64-70, 2012. Disponible en: https://goo.gl/8Klti8.

[25] D. Jurafsky y J.H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2a ed.: Pearson Prentice Hall, 2009.

[26] (2016) *VoxForge*. [En línea]. Disponible en: http://www.voxforge.org

[27] X. Huang, A. Acero y H. Hon, *Spoken Language Processing, a guide to theory, algorithm and system development*, Prentice Hall, 2001.

[28] J.V. Peña, "Contribuciones al reconocimiento robusto de habla," tesis doctoral, Dpto. de Teoría de la Señal y Comunicaciones, UC3M, Madrid, España, 2007. [En línea]. Disponible en: https://goo.gl/raEq5L

[29] F.J. Hernando Pericas, "Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos," tesis doctoral, Dpto. de Teoría de la Señal y Comunicaciones, UPC, Barcelona, España, 1993.

[30] Microsoft (2016) *Microsoft Developer Network. System.Speech Programming Guide for.NET Framework.* [En línea]. Disponible en: https://goo.gl/PM20D6.

[31] G.E Dahl et al, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, num. 1, pp. 30-42, ene. 2012. DOI: 10.1109/TASL.2011 .2134090

[32] Microsoft (2016) *Microsoft Developer Network Introducing Computer Speech Technology. Speech Server 2004 R2*. [En línea]. Disponible en: http://msdn.microsoft.com/en-us/library/ms870025

[33] *Guía y procedimiento de medida del ruido de actividades en el interior de edificios. Según anexo IV del Real Decreto 1367/2007*, AECOR, España, 2011. [En línea]. Disponible en: https://goo.gl /ra4EHQ